18

Missing Data and Dropout: Multiple Imputation and Weighting Methods

18.1 INTRODUCTION

In the previous chapter we distinguished between different types of missing data mechanisms by their assumptions concerning whether missingness is related to observed and unobserved responses. We emphasized that conventional likelihood-based analyses of incomplete data (e.g., linear mixed effects models fitted using standard statistical software such as PROC MIXED in SAS, the `lme` function in the `nlme` package in R and S-Plus, and the `xtmixed` command in Stata) yield valid inferences when data are MAR if the joint distribution of the vector of responses has been correctly specified. When the joint distribution is assumed to be multivariate normal, this requires correct specification of not only the model for the mean response, but also the model for the covariance among the responses. In this chapter we discuss two alternative approaches for handling missing data: multiple imputation and weighting methods. Both approaches are appealing in settings where a conventional likelihood-based analysis is no longer straightforward. For example, when there are missing covariates as well as missing responses, a likelihood-based analysis is not straightforward and cannot be implemented using standard statistical software. Also there is no convenient specification of the joint distribution of the vector of responses for marginal models when the responses are discrete; instead, the generalized estimating equations (GEE) approach is routinely used and requires the stronger assumption of MCAR. In both of these settings, as well as others, multiple imputation and weighting methods provide a convenient and practical method for handling missing data in longitudinal analyses.

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

18.2 MULTIPLE IMPUTATION

Multiple imputation is a flexible method for handling missing data that has recently been implemented in numerous commercially available software packages (e.g., SAS, Stata, SPSS, R, and S-Plus), as well as in more specialized software programs (e.g., SUDAAN and Solas). As discussed in Chapter 17, imputation is an intuitively simple technique: we "fill in" or impute plausible values for the missing data, thereby creating a "completed" data set that can be analyzed using standard statistical methods for complete data. Imputation allows us to proceed with the analysis of the completed data set as though there were no missing data at all. However, if each missing datum is filled in with one plausible value only, any subsequent analysis of this single completed data set is problematic. The trouble with such an analysis is that the imputed values are implicitly treated as though they are known, neglecting the fact that there is inherent uncertainty surrounding the imputed values. Conventional methods for standard error estimation do not properly account for this uncertainty. Specifically, any analysis of the single imputed data set as though it were a complete data set will, in general, produce anti-conservative results. By "anti-conservative," we mean that nominal p-values will tend to be too small and confidence intervals will be too narrow. Multiple imputation was developed to correct this problem.

With multiple imputation methods each missing datum is filled in with plausible values multiple times, producing multiple completed data sets. The replacement of missing data with multiple plausible values ensures that the uncertainty associated with the imputed values can be properly accounted for. Typically the number of imputations is relatively small, say between 10 and 25. Each of the completed data sets is then analyzed using standard methods for complete data, as if there were no missing data. These analyses produce a set of results that 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 missing data. Specifically, if we assume that $m > 1$ filled-in data sets are created, then m different estimates of the regression parameters β, say $\widehat{\beta}^{(k)}$ (for $k = 1, ..., m$), can be obtained from the separate analyses of each of the m data sets. In addition the m analyses of the filled-in data sets also yield m estimates of the covariance of $\widehat{\beta}^{(k)}$, for $k = 1, ..., m$. The multiple imputation estimate of β is simply the unweighted average of the m estimates,

$$\widehat{\beta} = \overline{\beta} = \frac{1}{m} \sum_{k=1}^{m} \widehat{\beta}^{(k)}.$$

The estimated covariance of $\widehat{\beta}$ is given by

$$\widehat{\mathrm{Cov}}(\widehat{\beta}) = W + (1 + m^{-1})B,$$

where

$$W = \frac{1}{m} \sum_{k=1}^{m} \widehat{\mathrm{Cov}}(\widehat{\beta}^{(k)})$$

and

$$B = \frac{1}{m-1}\sum_{k=1}^{m}\left(\widehat{\beta}^{(k)} - \overline{\beta}\right)\left(\widehat{\beta}^{(k)} - \overline{\beta}\right)'.$$

Although the expression for $\widehat{\mathrm{Cov}}(\widehat{\beta})$ appears somewhat complicated, it simply combines two inherent sources of variability: the within-imputation variability (W) and the between-imputation variability (B).

So far, we have briefly described how to make inferences from a statistical analysis based on multiple imputation. It can be thought of as a three-step process. First, the missing data are filled in m times to create m completed data sets. Second, the m completed data sets are analyzed using standard statistical methods (e.g., fitting any of the regression models discussed in Parts II and III). Finally, the results from the m analyses of the completed data sets are combined in the manner described above. For the remainder of this section, we focus on the first step and consider how to create multiple completed data sets using methods that incorporate appropriate variability across the m imputations. At the outset it must be recognized that there are numerous different methods of imputation, although most have a similar basis. Moreover the method of choice is often determined by the patterns of missingness (e.g., monotone versus non-monotone patterns of missingness) in the data set at hand. Although there are numerous methods of imputation, one general principle should guide the choice of imputation method: "proper" imputations should be drawn at random from the conditional (or so-called predictive) distribution of the missing data given the observed data. In general, a proper imputation of Y_i^M is obtained by *randomly* drawing values from $f(Y_i^M|Y_i^O, X_i)$. By choosing to draw values from $f(Y_i^M|Y_i^O, X_i)$, we are implicitly assuming that missingness is MAR; that is, the predictive distribution of the missing data, given the observed data, does not depend on the observed response pattern, R_i, with $f(Y_i^M|Y_i^O, X_i, R_i) = f(Y_i^M|Y_i^O, X_i)$. To randomly sample values from $f(Y_i^M|Y_i^O, X_i)$, it is important to distinguish two settings: (1) monotone missing data patterns, and (2) non-monotone missing data patterns. Imputation is far more straightforward in the former case, and much of the remainder of this section focuses on monotone missing data. For the latter case, iterative computational methods are usually required.

Before describing specific methods for imputing longitudinal data with missing responses, it is worth noting that these methods implicitly assume the data set is structured in a "wide" rather than a "long" format, with a single "record" for each individual. In a "wide" format, methods for imputation of a missing response at any particular occasion can exploit the positive correlation with the responses at any of the remaining occasions. This is discussed in greater detail in Section 18.6.

18.2.1 Monotone Missing Data Patterns

As noted in the previous chapter, monotone missing data patterns arise in longitudinal studies when missingness occurs only through dropout. For example, suppose that the first response, Y_1, is always observed but subsequent responses are missing due to dropout. Then a monotone missing pattern is produced where Y_1 is fully observed,

the second response, Y_2, has the fewest missing values, Y_3 has the second fewest missing values, and so on. With monotone missing data patterns, missing values can be imputed by first fitting an appropriate model (e.g., a regression model) to predict Y_{i2} from Y_{i1} and X_i and then randomly sampling from this model to impute the missing values in the second response. Next the missing values in the third response can be imputed based on an appropriate model for predicting Y_{i3} from Y_{i1} and both the *observed* and *imputed* values of Y_{i2}. Imputation of the remaining missing values can continue in a similar way until all of the missing values have been filled in. The resulting set of imputed values is a proper imputation of Y_i^M from $f(Y_i^M|Y_i^O, X_i)$ when the MAR assumption holds. There are two commonly used methods of imputation for monotone missing data patterns: regression methods and predictive mean matching methods. We briefly describe each approach in turn.

Regression Methods

In regression methods for imputing longitudinal data, the missing responses are imputed sequentially using all preceding responses in the monotone pattern (and any subset of the covariates) as "predictors" in a regression model. That is, a series of regression models for Y_{ik}, given $Y_{i1}, \ldots, Y_{ik-1}$ and X_i (or a subset of X_i), are fitted to the observed data. When the responses are continuous, standard linear regression is widely used to generate imputations. For example, when Y_{ik} is continuous, a linear regression model

$$E(Y_{ik}|Y_{i1}, \ldots, Y_{ik-1}, X_i) = \gamma_1 + \gamma_2 Y_{i1} + \cdots + \gamma_k Y_{ik-1},$$

can be fitted using the observed data on subjects who have not dropped out by the k^{th} occasion; alternative models may be needed when linearity is insufficient to capture the functional forms for the relationships. For simplicity, we have assumed no dependence on X_i in the regression model above. More generally, a series of linear regression models

$$E(Y_{ik}|Y_{i1}, \ldots, Y_{ik-1}, X_i) = Z_{ik}'\gamma,$$

can be fitted to data on the N_k subjects who have not dropped out by the k^{th} occasion. In a departure from the notation used in previous chapters, we let Z_{ik} denote a vector comprised of $Y_{i1}, \ldots, Y_{ik-1}$ and any subset of the components of X_i. In this model γ denotes a $q \times 1$ vector of regression parameters relating Y_{ik} to the preceding responses and covariates. Also, in a slight abuse of notation, note that there is a separate set of regression parameters, γ, for the regression model at each occasion and that the corresponding dimensions of Z_{ik} and γ can vary in the models for different occasions. The linear regression models introduced here, however, are for the purpose of imputation and are not the same models that were discussed in Part II for longitudinal analyses. The regression models considered here are imputation and not analysis models.

For the regression model at the k^{th} occasion,

$$E(Y_{ik}|Y_{i1}, \ldots, Y_{ik-1}, X_i) = Z_{ik}'\gamma,$$

the regression parameters (and the residual variance) can be estimated via ordinary least squares (OLS). The fitted linear regression produces estimates of the regression parameters, $\widehat{\gamma}$, and their associated covariance matrix,

$$\widehat{\mathrm{Cov}}(\widehat{\gamma}) = \widehat{\sigma}^2 \left(\sum_{i=1}^{N_k} Z_{ik} Z_{ik}' \right)^{-1},$$

where $\widehat{\sigma}^2$ is an estimate of the residual variance and Z_{ik} is the design vector for the regression of Y_{ik} on $Y_{i1}, \ldots, Y_{ik-1}$, and any subset of the components of X_i.

In principle, the fitted regression could be used to produce predictions or imputations of the missing values of Y_{ik}. However, because the fitted regression produces a deterministic prediction of the missing values on Y_{ik} for any fixed Z_{ik}, we must incorporate random variation to reflect the uncertainty of the imputations. Specifically, we need to add to the predicted value for Y_{ik} a random draw from the residual distribution of Y_{ik} given $Y_{i1}, \ldots, Y_{ik-1}$ and X_i. This corresponds to adding a random "error" from the regression model. In addition we must account for one more source of uncertainty. Recall that the prediction of Y_{ik} is based on a set of *estimated* regression coefficients and an estimate of the residual variance, σ^2; however, implicitly, the imputation process above would be treating these as fixed and known rather than sample estimates. A "proper" imputation should also account for this latter source of variability. To incorporate this additional source of variation, each imputation should be based on different (i.e., randomly perturbed) values for the regression coefficients and σ^2. (Specifically, these values should be random draws from what is known as their "posterior distributions". A detailed description of "posterior distributions" is omitted because it requires some basic understanding of Bayesian statistics, a topic that is beyond the scope of this chapter). This last step, randomly drawing from the "posterior distribution" of the parameters (under a so-called non-informative prior distribution for these parameters), is probably the most complicated part of the imputation process. Fortunately, this step has been implemented in many statistical software packages for multiple imputation (e.g., PROC MI in SAS assuming multivariate normality).

To summarize, regression methods require the use of the following two steps to produce imputed values, say Y_{ik}^*, for the missing Y_{ik}:

1. New regression parameters, say γ^*, and the residual variance, say σ^{*2}, are randomly drawn from their "posterior distributions" to account for the uncertainty in estimating γ and σ^2. Specifically, the residual variance is randomly drawn as

$$\sigma^{*2} = (N_k - q)\widehat{\sigma}^2 / \chi^2,$$

where $N_k - q$ denotes the degrees of freedom for the residual variance, and χ^2 is a random draw from a chi-square distribution with $(N_k - q)$ degrees of freedom. Then the regression parameters, γ^*, are randomly drawn from a multivariate normal distribution with mean equal to the estimated regression

parameters, $\widehat{\gamma}$, and with covariance matrix,

$$\widehat{\mathrm{Cov}}(\widehat{\gamma}) = \sigma^{*2}\left(\sum_{i=1}^{N_k} Z_{ik}Z_{ik}'\right)^{-1},$$

where $\widehat{\sigma}^2$ has been replaced by σ^{*2} and Z_{ik} denotes the design vector for the regression of Y_{ik} on $Y_{i1}, \ldots, Y_{ik-1}$, and any subset of the components of X_i.

2. The missing values for Y_{ik}, say Y_{ik}^*, can then be imputed on the basis of the following predictions:

$$Y_{ik}^* = Z_{ik}'\gamma^* + e^*,$$

where, for each missing observation on Y_{ik}, e^* is randomly drawn from a normal distribution with mean zero and standard deviation, σ^*.

Regression imputation of the remaining missing values continues in a similar manner until all of the missing values have been filled in. Multiple imputations are obtained by repeating these steps $m > 1$ times.

The description of regression imputation given above has focused on the case where the longitudinal responses are continuous. When the responses are discrete rather than continuous (e.g., repeated binary responses), a regression imputation can be based on a series of suitable generalized linear models,

$$g\{E(Y_{ik}|Y_{i1}, \ldots, Y_{ik-1}, X_i)\} = Z_{ik}'\gamma,$$

where $g(\cdot)$ is a known link function. For example, with binary responses, the missing responses can be imputed sequentially using all preceding responses in the monotone pattern (and any subset of the covariates) as predictors in a series of logistic regression models. The logistic regression model at the k^{th} occasion,

$$\mathrm{logit}\{\Pr(Y_{ik} = 1|Y_{i1}, \ldots, Y_{ik-1}, X_i)\} = Z_{ik}'\gamma,$$

can be estimated using standard statistical software for logistic regression. Based on the estimated parameters for the logistic regression model, say $\widehat{\gamma}$, new logistic regression models are obtained by randomly drawing logistic regression parameters, say γ^*, from their "posterior distribution." The missing binary responses can then be imputed from Bernoulli distributions with probabilities of success,

$$\frac{\exp(Z_{ik}'\gamma^*)}{1 + \exp(Z_{ik}'\gamma^*)},$$

determined by the randomly drawn logistic regression parameters. In a similar way, when the longitudinal responses are counts, regression imputation can be based on loglinear regression models where missing responses are imputed from Poisson distributions.

Predictive Mean Matching

The second approach to imputation is known as predictive mean matching and is closely related to regression methods. Predictive mean matching is also based on a sequence of regression models for Y_{ik}, given $Y_{i1}, \ldots, Y_{ik-1}$ and X_i, fit to the observed data. However, instead of filling-in the missing values on the basis of predicted values from the regressions, predictive mean matching imputes using the observed values of the outcomes from the data at hand that are in a certain sense "closest" to the predicted values.

As with regression imputation methods, a series of regression models for Y_{ik}, given $Y_{i1}, \ldots, Y_{ik-1}$ and X_i, are fit using the observed data. When the responses are continuous, standard linear regression is typically used to obtain predicted values. The fitted linear regression produces estimates of both the regression parameters and the residual variance. New regression parameters and residual variance are then randomly drawn from their "posterior distributions" to account for the uncertainty in their estimation. Given a random draw of γ from its "posterior distribution", say γ^*, for each missing value a predicted value

$$\widehat{Y}_{ik} = Z'_{ik}\gamma^*$$

is calculated. Note that unlike in regression imputation where a random "error" is also included in the prediction, here $\widehat{Y}_{ik}$ is a prediction of the mean. Predicted values are also calculated for observations with non-missing values for Y_{ik}. From the latter, a subset of K observations whose corresponding predicted values are closest to $\widehat{Y}_{ik}$ is generated; the predictive mean matching method requires the number of closest observations, K, to be specified in advance. Thus for each missing value there is a set of K potential "donors" who have similar predicted values. The missing value is then replaced by a value drawn randomly from these K observed values. Multiple imputations are obtained by repeating these steps $m > 1$ times.

Although predictive mean matching is closely related to regression methods, it should be somewhat less sensitive to any misspecification of the sequence of regression models. That is, predictive mean matching relies on the regression models only to determine the distance between observed and missing values. Instead of replacing the missing values using predictions from the regression models, predictive mean matching imputes using the observed values of the outcomes that are "closest" to the predicted values. Of note, because imputed values are drawn randomly from observed values contributed from other subjects, the predictive mean matching method ensures that all imputed values are plausible.

Finally, we note that there is a third approach to imputation known as "propensity score" methods. In propensity score methods, values to impute for the missing responses are randomly drawn from observations on subjects who are equally likely to dropout but do not at that occasion. Propensity score methods require a model for the probability of dropout at any occasion; ordinarily it is assumed that dropout depends only on past observed responses and any subset of the observed covariates (i.e., the dropout mechanism is MAR). However, a potential drawback of this approach is that the resulting set of imputed values is not, in general, a proper imputation of

Y_i^M from $f(Y_i^M|Y_i^O, X_i)$. As a result this method of imputation may not preserve the inter-relationships among the variables.

18.2.2 Non-monotone Missing Data Patterns

When the missing data patterns are monotone it greatly simplifies the process of drawing "proper" imputations. There are many methods of imputation with monotone missingness; in the previous section we have focused on two commonly used methods. When the missing data patterns are non-monotone or intermittent, iterative and more computationally demanding methods are usually required because it is no longer straightforward to randomly sample from $f(Y_i^M|Y_i^O, X_i)$. In this section we focus on two commonly used algorithms for simulating random draws from $f(Y_i^M|Y_i^O, X_i)$. The first algorithm, known as "data augmentation" (DA), yields a random sample from $f(Y_i^M|Y_i^O, X_i)$ after a sufficiently large number of iterations of the algorithm. The second algorithm, known as "chained equations," is somewhat ad hoc and produces imputations that are, at best, a random sample from an approximation to $f(Y_i^M|Y_i^O, X_i)$. Despite the fact that the "chained equations" approach rests on shaky statistical foundations, it appears to work reasonably well in practice.

Data augmentation is best understood as an iterative two-step algorithm that alternates between random draws of the missing responses, given current values of the imputation model parameters, and random draws of the model parameters, given both the observed and current imputed values of the responses. Specifically, DA involves iterating between the following two steps:

1. randomly sampling the missing responses, Y_i^M, from $f(Y_i^M|Y_i^O, X_i)$ given a current random draw of the imputation model parameters, and

2. randomly sampling the imputation model parameters given a current random draw of the missing data.

The resulting sequence of draws of the missing data, Y_i^M, and the imputation model parameters, form what is known as a Markov chain. Note that imputations of missing responses in step 1 are not necessarily draws from the true $f(Y_i^M|Y_i^O, X_i)$ because they are based on imputation model parameters drawn (in step 2) from the distribution that results from treating the imputed values as if they were actual observed values. However, after a sufficiently large number of iterations of this two-step algorithm, the components of this Markov chain have the desired distribution and Y_i^M in step 1 can be considered a random sample from $f(Y_i^M|Y_i^O, X_i)$. Data augmentation is often referred to as a Markov chain Monte Carlo (MCMC) method because it involves repeated random sampling (also known as Monte Carlo simulation) of a Markov chain whose distribution, after a sufficiently large number of iterations, converges or stabilizes to $f(Y_i^M|Y_i^O, X_i)$.

When it is assumed that the responses have a multivariate normal distribution, the first step of the DA algorithm is relatively straightforward because $f(Y_i^M|Y_i^O, X_i)$ also has a (multivariate) normal distribution. That is, given some current values of the mean vector and the covariance matrix, imputing the missing responses only requires

drawing a random sample from a conditional (multivariate) normal distribution. In contrast, the random sampling in the second step is somewhat more involved. In the second step, given the observed and current imputed values for the responses (i.e., given a so-called completed data set), new model parameters must be obtained by sampling from the "posterior distribution" of the mean vector and covariance matrix. Without any prior information about the mean vector and covariance matrix, these too can be simulated from well-known distributions. Specifically, the mean vector can be randomly sampled from a multivariate normal distribution and the covariance matrix can be randomly sampled from an inverted Wishart distribution; a more detailed description of sampling from the posterior distribution of the mean vector and covariance matrix is outside the scope of this chapter. These new parameter values are then used in the first step and the process iterates, creating a Markov chain. Given a sufficiently large number of iterations, the imputed values for Y_i^M in step 1 can be considered a random sample from $f(Y_i^M|Y_i^O, X_i)$.

The use of MCMC methods for generating imputations requires some additional care because (1) the Markov chain may require many iterations before the desired stationary distribution is obtained and (2) the Markov chain has an inherent dependence, in the sense that the current state of the chain has some influence on the next state in the iteration (i.e., with any Markov chain, the "current" state is predictive of the "future"). To address the first concern, a large number, say 1000 to 5000, of initial or "burn-in" iterations of the Markov chain are run, from which no imputations are made. In a sense, samples from a large number of iterations at the beginning of the algorithm are simply ignored; these burn-in iterations are executed before the first imputation is drawn from the Markov chain. The large number of "burn-in" iterations used at the start of the Markov chain increases the likelihood that the desired stationary distribution that we wish to sample from has been achieved. This also removes any dependence on the starting value selected for the chain (ordinarily determined by the computer's random seed). To address the second concern about dependence, successive iterations of the chain are avoided as they tend to be correlated. Instead, the chain is subsampled, with imputations drawn from every k^{th} iteration of the chain (where $k > 1$). By choosing a relatively large value for k, say 100 to 500, any dependence between consecutive imputations drawn from the chain is negligible.

Next we consider a second iterative method for drawing imputations when the missing data patterns are non-monotone. This alternative method does not rely on the assumption that the responses have a multivariate normal distribution. The method is referred to as "multivariate imputation by chained equations" (MICE) and requires that a sequence of separate regression models be specified for each response with missing data.[1] The specific type of regression model selected depends on the type of response variable; for example, a linear regression model is commonly used for a quantitative response, a logistic regression model for a binary response, and a Poisson

[1]Implementations of the chained equation approach are available in the MICE library for R and S-Plus, ICE for Stata, and the IVEware macro for SAS (or standalone); a useful guide to software implementations is available at `http://multiple-imputation.com`.

regression model for counts. A notable feature of each of these regression models is that the response at any occasion is regressed on all other responses, both past and future responses (and any subset of the covariates). That is, MICE methods specify a sequence of regression models for $f(Y_{ij}|Y_{i1}, ..., Y_{i,j-1}, Y_{i,j+1}, ..., Y_{in}, X_i)$.

Once a sequence of regression models has been specified for the conditional mean of the response at each occasion, $E(Y_{ij}|Y_{i1}, ..., Y_{i,j-1}, Y_{i,j+1}, ..., Y_{in}, X_i)$, imputation by chained equations involves cycling through the fitting of these regression models, usually in order of the response variables that have the least amount of missing data. In each of these regression models, the response at any occasion is regressed on all other responses, both past and future (and any subset of the covariates), where any missing values among the "predictors" in the regression have been replaced by their imputed values from the previous cycle of the algorithm. After a particular regression model has been fit, imputed values for the missing responses can be generated from the fitted regression model, following the usual steps for a regression imputation to properly account for uncertainty. This sequence of imputing missing values for each response can be continued from one cycle to another, each time overwriting previously drawn values with updated imputed values. Typically this process continues iteratively for a pre-determined number of cycles (akin to "burn-in" iterations) and the set of imputations in the last cycle is used to generate a "completed" data set. Multiple imputations are obtained by repeating these steps $m > 1$ times.

One of the appealing features of imputation by chained equations is that it avoids the problem of implicitly assuming a multivariate normal model for the responses. By specifying imputation models on a variable by variable basis, different imputation models can be used for different types of responses; e.g., a logistic regression imputation model can be used for binary responses. However, the method is somewhat ad hoc and does not have a firm theoretical basis. Specifically, the set of conditional distributions specified by each regression model, $f(Y_{ij}|Y_{i1}, ..., Y_{i,j-1}, Y_{i,j+1}, ..., Y_{in}, X_i)$, may not even be compatible with any joint distribution for the vector of responses, $f(Y_{i1}, ..., Y_{in}|X_i)$. The potential consequences of this incompatibility on the quality of the imputations from MICE are unclear. So, at best, the imputations arising from this method should be regarded as random draws from an approximation to $f(Y_i^M|Y_i^O, X_i)$. With this caveat, there are results from limited simulation studies that suggest the method yields approximately valid inferences.

18.2.3 Concluding Remarks

Multiple imputation is a flexible method for handling missing data in longitudinal studies when the missing data mechanism is thought to be MAR. We note that many of the imputation models that have been reviewed in the previous sections assume a multivariate normal distribution for the response vector. Naturally, imputations based on a multivariate normal distribution for the responses will be far more appealing for the linear models for longitudinal data described in Part II of the book. For these linear models, recall that a likelihood-based analysis of the incomplete data is also valid under MAR. However, such an analysis does require that the likelihood be correctly specified. For example, maximum likelihood estimation of the regression parame-

ters in these linear models is valid when data are MAR provided that the assumed multivariate normal distribution for the responses has been correctly specified.

When the assumptions for a likelihood-based analysis hold, multiple imputation may offer few potential benefits. In settings where the assumptions (e.g., multivariate normality) underlying both approaches are tenable, the results should be relatively similar. In such settings ML estimation based on the incomplete data may be preferred on grounds that (1) it is less computationally intensive, (2) it yields a unique set of results, and (3) it provides the most efficient estimates of the model parameters. However, conditions 2 and 3 become less important as the number of imputed values m increases. That is, as m increases (or, more formally, as $m \longrightarrow \infty$), multiple imputation yields almost efficient estimates when compared to ML estimation with the incomplete data; in addition, as m increases, the final estimates yielded by multiple imputation become far more stable. Of course, when m is relatively large, the computational burden of multiple imputation is far greater relative to a likelihood-based analysis.

When does multiple imputation offer distinct advantages for the analysis of longitudinal studies with missing data? There are at least three scenarios where multiple imputation may be considered advantageous. First, when there are extraneous covariates that are thought to be either predictive of the probability of missingness and/or predictive of the responses. By extraneous, we mean covariates that would not ordinarily be included in the analysis model. These extraneous covariates are sometimes referred to as auxiliary variables because there is no interest in making inferences about them (or conditional upon them). In multiple imputation, these extraneous covariates can be introduced in the imputation process in a relatively straightforward way to potentially improve the imputation of missing values. For example, in a clinical trial, missingness may be related to side effects of the treatments. In such a setting, side effects is an extraneous covariate in the sense that it would not ordinarily be included in the analysis model. That is, ordinarily there is no scientific interest in an analysis of change in the response that conditions or stratifies on whether an individual experiences side effects. However, although side effects is considered an extraneous covariate for the analysis model, it can be incorporated into the imputation model. Indeed, the inclusion of any extraneous covariate that is highly correlated with the response is likely to improve the imputations. Of course, it should be acknowledged that inclusion of these extraneous covariates in the imputation model, but not in the analysis model, implies that there is some incompatibility between the two models. In ideal circumstances the two models, the imputation model and the analysis model, should agree in terms of their representation of the relationships among the variables. However, in practice, when the analysis model is simpler than the imputation model (e.g., the analysis model implicitly assumes no dependence between the responses and extraneous covariates), this type of incompatibility should not be of great concern.

The second scenario is when there are missing covariates in addition to missing responses. Likelihood-based analysis with incomplete covariates and responses is not straightforward and has not been implemented in standard statistical software. On the other hand, multiple imputation of both missing responses and covariates is,

in principle, a relatively straightforward way to handle this problem (albeit requiring some assumptions about the missing covariates).

Third, multiple imputation may be appealing in settings where a full likelihood-based analysis is not possible because there is no convenient specification of the joint multivariate distribution of the vector of responses. For example, likelihood-based analysis of marginal models is not at all straightforward when the vector of responses is discrete rather than continuous. In such settings the generalized estimating equations (GEE) approach is routinely used as an alternative method of estimation. Standard applications of the GEE approach are valid only under MCAR. However, certain multiple imputation methods (e.g., logistic regression methods) can be fruitfully combined with standard GEE analyses of discrete longitudinal responses to make it valid under MAR. Finally, we note that multiple imputation also provides a relatively flexible and general framework for undertaking sensitivity analyses. By considering a series of alternative imputation models for the missing data, the impact of variations in the imputation model on the overall results provides an assessment of their robustness.

18.3 INVERSE PROBABILITY WEIGHTED METHODS

In the previous section we discussed multiple imputation methods that replace missing values with randomly drawn values from the conditional distribution of the missing data given the observed data, denoted $f(Y_i^M|Y_i^O, X_i)$. In this section we consider an alternative approach for handling missing data that does not require any assumptions about $f(Y_i^M|Y_i^O, X_i)$; instead, an adjustment to the analysis is made by weighting 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. These weighting approaches are often called propensity weighted or inverse probability weighted (IPW) methods. In general, inverse probability weighted methods are more straightforward to implement when any missingness is restricted to dropout. In addition IPW methods are more appealing in settings where a full likelihood-based analysis is not possible due to the lack of a convenient specification of the joint multivariate distribution of the vector of responses, such as in settings where the vector of responses is binary rather than continuous. As a result the following description of IPW methods focuses exclusively on the problem of handling dropout in GEE analyses of discrete longitudinal responses.

Recall that when missingness is restricted to dropout, we can replace the vector of response indicators, $R_i = (R_{i1}, ..., R_{in})'$, by a scalar variable D_i, with $D_i = 1 + \sum_{j=1}^{n} R_{ij}$ denoting the occasion at which dropout occurs. For a so-called complete-case $Y_i = (Y_{i1}, ..., Y_{in})'$ and $D_i = n + 1$, whereas for an individual with an incomplete vector of n_i responses (where $n_i < n$), with observed components $Y_i^o = (Y_{i1}, ..., Y_{in_i})'$, $D_i = n_i + 1$.

The basic idea underlying all IPW methods is to base estimation on the observed responses but weight them to account for the inverse probability of remaining in the study. The propensities for dropout (or, conversely, for subjects 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 (e.g., subject characteristics) that are thought likely to predict dropout. To provide some intuition for IPW methods, we first consider the simplest version of this approach based on an adjustment to the complete-case analysis. For an IPW complete-case analysis we need to estimate a single weight, say w_i, only for those subjects who complete the study ($D_i = n + 1$). The weight w_i can be computed as the inverse of the product of the conditional probabilities of remaining in the study at each occasion,

$$w_i = (\pi_{i1} \times \pi_{i2} \times \cdots \times \pi_{in})^{-1},$$

where $\pi_{ij} = \Pr(D_i > j | D_i \geq j) = \Pr(R_{ij} = 1 | R_{i1} = \cdots = R_{i,j-1} = 1)$. For the first occasion, it is usually assumed that $R_{i1} = 1$ for all individuals, and then $\pi_{i1} = 1$. For all subsequent occasions, the $\pi_{ij}{}'s$ can be estimated from those individuals remaining in the study at the $(j-1)^{th}$ occasion, given their recorded history of observed responses and covariates up to the $(j-1)^{th}$ occasion. The weight w_i is then estimated as the inverse of the product of the estimated $\pi_{ij}{}'s$. In an IPW complete-case analysis, each contribution to the analysis from the "completers" receives a weight of w_i. The intuition behind this weighting 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 but who did not complete the study. For example, a subject with weight of 4 has a probability of completing the study of 0.25 (or $\frac{1}{w_i} = 0.25$). As a result, in a complete-case analysis, data from this subject should count once for herself and 3 times for those subjects who do not complete the study (recall that if the probability of completing the study is $\frac{1}{4}$, it means that 3 subjects are expected to dropout for every one that completes the study). Given $\widehat{w}_i$, the inverse of the estimated probability that the i^{th} subject completes the study, an IPW complete-case analysis can then be performed. For example, the standard GEE approach can be readily adapted to handle dropout that is MAR by making an appropriate adjustment to the analysis for the propensity for dropout. Specifically, an inverse probability weighted GEE (IPW-GEE) can be used to analyze the data from the "completers" only, where each subject's contribution to the analysis is weighted by $\widehat{w}_i$.

An IPW complete-case analysis is valid provided that the model that produces the estimated w_i is correctly specified. The w_i are not ordinarily known when there is dropout in longitudinal studies, but must be estimated from the observed data (e.g., using a repeated sequence of logistic regressions to model the π_{ij}'s). Under the assumption that data are MAR, the π_{ij}'s are assumed to be a function only of the *observed* covariates and the *observed* responses prior to dropout. The π_{ij}'s can also depend on any additional variables or subject characteristics that are thought likely to predict dropout. Therefore estimation of the weights is, in principle, straightforward. However, an IPW analysis restricted to the complete-cases is less than optimal in the sense that it makes very inefficient use of the available data. Next we discuss how the IPW method for handling dropout can be made more efficient by conducting an appropriately weighted available-data analysis, with weights that are also occasion-

specific. Specifically, we focus on describing a general application of IPW methods to the GEE analysis of longitudinal data.

Recall from Chapter 13 that the standard GEE estimator is obtained as the solution to the following estimating equations:

$$\sum_{i=1}^{N} D_i^{o\prime} V_i^{o-1} \left(Y_i^o - \mu_i^o\right) = 0,$$

where $D_i^o = \frac{\partial \mu_i^o}{\partial \beta}$, and $\mu_i^o = g^{-1}(X_i^o \beta)$ denotes the components of μ_i corresponding to the observed components of the response vector $Y_i^o = (Y_{i1}, ..., Y_{in_i})'$. As was discussed in Chapter 13, the solution to these estimating equations yields a consistent estimator of β provided the data are MCAR (or provided that missingness depends only on the covariates included in the model for the mean response). However, when dropout is MAR, the standard GEE can yield badly biased estimates of β. The inverse probability weighted GEE (IPW-GEE) approach was developed to circumvent this specific problem. In the IPW-GEE the dropout process is accounted for by appropriately weighting the estimating equations. Specifically, the IPW-GEE estimator is obtained as the solution to the following *weighted* estimating equations:

$$\sum_{i=1}^{N} D_i' V_i^{-1} W_i \left(Y_i - \mu_i\right) = 0,$$

where D_i is the $n \times p$ derivative matrix, V_i is a $n \times n$ working covariance matrix for Y_i, and W_i is an $n \times n$ diagonal matrix of the occasion-specific weights, w_{ij}, for $j = 1, ..., n$,

$$W_i = \begin{pmatrix} R_{i1} \times w_{i1} & 0 & \cdots & 0 \\ 0 & R_{i2} \times w_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R_{in} \times w_{in} \end{pmatrix}.$$

Note that if the i^{th} individual's response is observed at the j^{th} occasion, it receives weight of w_{ij}; in contrast, all of the unobserved responses receive weight of zero. The weight, w_{ij}, is the inverse of the *unconditional* probability of being observed at the j^{th} occasion.

To calculate these weights, let π_{ij} denote the *conditional* probability of the i^{th} individual being observed (or not dropping out) at the j^{th} occasion, given that this individual was observed at the prior occasions. For the first occasion it is usually assumed that $R_{i1} = 1$ for all individuals, and then $\pi_{i1} = 1$. Recall that the MAR assumption implies that

$$\begin{aligned} \pi_{ij} &= \Pr(D_i > j | D_i \geq j, X_i, Y_i) \\ &= \Pr(R_{ij} = 1 | R_{i1} = \cdots = R_{i,j-1} = 1, X_i, Y_i) \\ &= \Pr(R_{ij} = 1 | R_{i1} = \cdots = R_{i,j-1} = 1, X_i, Y_{i1}, ..., Y_{i,j-1}), \end{aligned}$$

with π_{ij} depending only on fully-observed covariates and previously-observed responses. It is also possible to allow the π_{ij} to depend on additional fully-observed variables (e.g., subject characteristics) that are thought to be predictive of dropout. For the IPW-GEE analysis the required weight w_{ij} for the i^{th} individual at the j^{th} occasion is the inverse of the *unconditional* probability of being observed at the j^{th} occasion. The *unconditional* probability of being observed at the j^{th} occasion can be expressed as the cumulative product of conditional probabilities,

$$\pi_{i1} \times \pi_{i2} \times \cdots \times \pi_{ij}.$$

The required weight is then given by the inverse of the cumulative product of conditional probabilities,

$$w_{ij} = (\pi_{i1} \times \pi_{i2} \times \cdots \times \pi_{ij})^{-1}.$$

Because we assumed that $R_{i1} = 1$, the first diagonal element of W_i is fixed at 1, while the remaining elements have weights given by w_{ij} for occasions when the response is observed ($R_{ij} = 1$) and weights of zero for occasions when the response is not observed ($R_{ij} = 0$).

Two somewhat technical, yet important, details are worth mentioning at this stage. The first relates to the magnitudes of the estimated weights. With IPW methods, extra care needs to be exercised when certain configurations of values for the "covariates" (i.e., $X_i, Y_{i1}, ..., Y_{i,j-1}$) yield estimates of the π_{ij}'s that are very small. When the estimates of π_{ij} are very small and close to zero, they can lead to estimated weights that are extremely large. Analyses that incorporate such extreme weights will be unduly influenced by the small subset of observations with these weights, yielding estimators of the regression parameters that are unstable and have poor precision. We recommend examining the distribution of the estimated weights for the presence of discernibly large values and checking the sensitivity of results to the inclusion of observations that receive large weights. If the results of the analysis are quite sensitive to a small number of large weights, then the IPW analysis should be reconsidered in favor of alternative methods of adjusting for missingness.

The second issue relates to the choice of working covariance matrix in the IPW-GEE. Unless the working covariance matrix, V_i, is assumed to be diagonal, the IPW-GEE requires that the covariates, X_i, are fully-observed at all occasions. That is, the covariates are assumed to be known at both the occasions where the response is observed and those occasions where the response is unobserved. This will often be the case in designed studies, where the main components of X_i are treatment or exposure group indicators (i.e., time-invariant covariates), in addition to indicators of, or functions of, the intended times of measurement. However, in cases where not all components of X_{ij} are known when Y_{ij} is missing, the IPW-GEE estimator can be used if a "working independence" assumption is made. The "working independence" assumption corresponds to setting the off-diagonal elements of V_i to zero. When a "working independence" assumption is made, valid standard errors can be obtained by using the sandwich variance estimator. Furthermore we note that, when attempting to implement the IPW-GEE approach using standard statistical software for GEE, the "working independence" assumption may be required to ensure that the weights are

appropriately incorporated in the analysis. For example, occasion-specific weights can be specified by using the WEIGHT statement in PROC GENMOD in SAS or the `pweight` option for the `glm` command in Stata; however, these weights are appropriately incorporated in the IPW-GEE analysis only under a "working independence" assumption.

Thus an important property of the IPW-GEE is that the choice of working covariance matrix only has an impact on the efficiency of estimation. The IPW-GEE does not require correct specification of the working covariance matrix to consistently estimate the components of β and their standard errors. It does, however, require correct specification of the model for the dropout process, that is, for valid estimation of β, it requires that the model that produces the estimated weights be correctly specified.

Next we consider estimation of the weights. Recall that the MAR assumption implies that π_{ij} depends only on *observed* covariates and *observed* past responses. Therefore we can estimate π_{ij} (for $j > 1$) by, for example, constructing a logistic regression model for π_{ij},

$$\begin{aligned} \text{logit}(\pi_{ij}) &= \text{logit}\{\Pr(R_{ij} = 1 | R_{i1} = \cdots = R_{i,j-1} = 1, Z_{ij})\} \\ &= \theta_1 Z_{ij1} + \theta_2 Z_{ij2} + \cdots + \theta_q Z_{ijq} \\ &= Z_{ij}'\theta, \end{aligned}$$

where Z_{ij} is a $q \times 1$ design vector that incorporates certain components of X_{ij} and the past responses $(Y_{i1}, ..., Y_{i,j-1})$, and perhaps indicator or dummy variables for occasions. Note that Z_{ij} can also incorporate additional covariates that may be predictive of dropout but are not of subject-matter interest in the marginal model for the mean response; this is an especially appealing feature of IPW methods. Estimates of the logistic regression parameters, θ, can be obtained using standard statistical software for fitting logistic regression. That is, we can create a "stacked" data set in which each individual contributes a sequence of binary "outcomes" to the analysis, where each binary "outcome," R_{ij}, is an indicator of whether the response was observed at a given occasion, from the second occasion (because it is assumed that $R_{i1} = 1$) until either the occasion when dropout occurs or the last intended measurement occasion. Thus, individuals who do not dropout contribute a sequence of $n-1$ binary responses $(R_{i2}, \ldots, R_{in}$, where $R_{i2} = \cdots = R_{in} = 1)$ to the logistic regression analysis, whereas an individual who drops-out at the k^{th} occasion contributes $k-1$ binary responses $(R_{i2}, \ldots, R_{ik}$, where $R_{i2} = \cdots = R_{i,k-1} = 1, R_{ik} = 0)$. The covariates, Z_{ij}, in the logistic regression model can include certain components of X_{ij} and the previous observed responses $(Y_{i1}, ..., Y_{i,j-1})$, and perhaps indicator variables for occasions. The logistic regression analysis can also include additional covariates that are thought to be predictive of dropout but are not incorporated in X_{ij}. A standard logistic regression analysis of the "stacked" data set (often referred to as a "pooled" logistic regression) provides estimates of θ; the π_{ij}'s and the required weight w_{ij} can then be estimated as

$$\widehat{w}_{ij} = \left(\widehat{\pi}_{i1} \times \widehat{\pi}_{i2} \times \cdots \times \widehat{\pi}_{ij}\right)^{-1}.$$

Once the set of weights have been determined, they can be incorporated in a relatively straightforward way into standard statistical software for longitudinal analyses,

for example, using the WEIGHT statement within PROC GENMOD in SAS or the `pweight` option for the `glm` command in Stata. The resulting IPW-GEE estimates are unbiased for the regression parameters under a MAR dropout process, provided the model for estimating the dropout probabilities has been correctly specified. When the weights are incorporated in this manner, the standard errors are estimated by implicitly assuming the weights are fixed and known. In principle, it is possible to adjust the standard errors for the estimation of these weights. In practice, however, it is not straightforward to implement such an adjustment to the standard errors within existing statistical software for GEE. The formula for making an adjustment to the standard errors is somewhat involved; for completeness, it is outlined in a separate section at the end of this chapter that can be omitted at first reading without loss of continuity. Counter-intuitively, failure to account for the estimation of the weights will, in general, result in standard errors that are too large (i.e., estimation of the weights from the data at hand leads to improvements in precision). Therefore the unadjusted standard errors (e.g., based on the conventional sandwich variance estimator) provide valid inferences and can be considered to be slightly conservative. By "conservative", we mean that nominal p-values may be slightly larger and confidence intervals may be slightly wider than they should be. Until such time as the IPW-GEE with adjusted standard errors is implemented in widely available software, we recommend basing inferences on the unadjusted standard errors.

18.4 CASE STUDIES

In these case studies we illustrate some of the methods described earlier for handling missing data using two examples. The first example is based on data from the Treatment of Lead-Exposed Children (TLC) Trial (see Section 5.4). Recall that in this study the response variable, blood lead levels, is continuous and measured repeatedly at four occasions. Although the data on blood lead levels for the 100 children from the succimer and placebo groups are complete, for pedagogical purposes we created an incomplete data set with a non-monotone pattern of missing values generated under a MAR mechanism. The second example is from a longitudinal clinical trial of contracepting women (Machin et al., 1988) discussed in Sections 14.7 and 15.5. In this trial the response variable is binary, indicating whether a women experienced amenorrhea in four successive injection intervals. This trial had substantial dropout.

Treatment of Lead-Exposed Children (TLC) Trial

In our first illustration, we focus on methods for handling missing data on a continuous response. Recall that the TLC trial was a placebo-controlled, randomized trial of an orally administered chelating agent, succimer, in children with confirmed blood lead levels of 20 to 44 μg/dL. The following analyses are based on data on blood lead levels at baseline (or week 0), week 1, week 4, and week 6 for 100 children from the succimer and placebo groups of the TLC trial.

Table 18.1 Estimated regression coefficients and standard errors based on an analysis of response profiles of the blood lead level data at baseline, week 1, week 4, and week 6 for the children from the TLC trial.

Variable	Group	Week	Estimate	SE	Z
Intercept			26.272	0.710	36.99
Group	S		0.268	1.005	0.27
Week		1	−1.612	0.792	−2.04
Week		4	−2.202	0.815	−2.70
Week		6	−2.626	0.889	−2.96
Group × Week	S	1	−11.406	1.120	−10.18
Group × Week	S	4	−8.824	1.153	−7.66
Group × Week	S	6	−3.152	1.257	−2.51

In Chapter 5 the results of an analysis of response profiles of complete data on these 100 children were presented (see Section 5.4). The REML estimates of the regression parameters, and their standard errors, are reproduced in Table 18.1. For ease of interpretation, baseline (week 0) is chosen as the reference level for time and the placebo group is chosen as the reference level for treatment. In the TLC trial the question of main scientific interest concerns the comparison of the two treatment groups in terms of their mean changes from baseline. This question translates directly into a test of the three single-degree-of-freedom contrasts for the group × time interaction. The results in Table 18.1 indicate that children treated with succimer have a discernibly greater decrease in mean blood lead levels from baseline at all occasions when compared to the children treated with placebo. For example, when compared to the placebo group, the succimer group has an additional 3.152 μg/dL (with SE = 1.257) decrease in mean blood lead levels from baseline to week 6. There are even larger differences between the two treatment groups earlier in the trial.

To create an incomplete data set, we applied the following missing at random (MAR) mechanism to the responses at weeks 1, 4, and 6,

$$\text{logit}\{\Pr(R_{ij} = 1 | Y_{i1}, \text{Group}_i)\} = \theta_1 + \theta_2 Y_{i1} \times \text{Group}_i \times (j-1), \quad j = 2, 3, 4.$$

When $\theta_2 = 0$, this mechanism is MCAR with a constant probability of missingness, $\frac{1}{1+\exp(\theta_1)}$. However, when $\theta_2 \neq 0$, missingness is allowed to depend on the baseline response (Y_{i1}) for children randomized to the succimer group ($\text{Group}_i = 1$), with the strength of that dependence increasing over time. Note that this missing data mechanism yields a non-monotone pattern of missingness. To generate an incomplete data set, we fixed $\theta_1 = 2.5$ and $\theta_2 = -0.03$; a negative value for θ_2 implies that children in the succimer group with higher blood lead levels at baseline have a greater

Table 18.2 Missing data patterns generated in the succimer and placebo groups from the TLC trial.

Group	Week 0	Week 1	Week 4	Week 6	Frequency	Percent
Succimer	O	O	O	O	18	36%
	O	O	O	M	12	24%
	O	O	M	O	5	10%
	O	O	M	M	6	12%
	O	M	O	O	2	4%
	O	M	O	M	3	6%
	O	M	M	O	1	2%
	O	M	M	M	3	6%
Placebo	O	O	O	O	42	84%
	O	O	O	M	2	4%
	O	O	M	O	3	6%
	O	M	O	O	2	4%
	O	M	M	O	1	2%

Note: O denotes observed response, M denotes missing response.

Table 18.3 Mean blood lead levels at baseline, week 1, week 4, and week 6 for the incomplete data on children from the TLC trial. Mean blood lead levels for the complete data are reported in parentheses.

Group	Baseline	Week 1	Week 4	Week 6
Succimer	26.5	13.4	15.2	19.3
	(26.5)	(13.5)	(15.5)	(20.8)
Placebo	26.3	24.5	24.1	23.7
	(26.3)	(24.7)	(24.1)	(23.6)

probability of being missing at subsequent occasions. This yielded the non-monotone patterns of missingness displayed in Table 18.2. In the placebo group, 6% of children have missing responses at week 1, 8% at week 4, and 4% at week 6; in the succimer group, the corresponding rates are 18% at week 1, 30% at week 4, and 48% at week 6. Because missingness is related to higher baseline blood lead levels in the succimer group, we might expect the mean blood lead levels in the succimer group to be lower, when compared to the means in the complete data set, at subsequent occasions. The discrepancy between the means in the complete and incomplete data sets should be most pronounced at week 6 when the rate of missingness in the succimer group is

Table 18.4 Estimated regression coefficients (and standard errors) based on an analysis of response profiles of (i) the complete data, denoted ML(C), (ii) the incomplete data, under the assumption of an unstructured covariance, denoted ML(UN), (iii) the incomplete data, under the (incorrect) assumption of independent responses and constant variance, denoted ML(IND), and (iv) multiple imputed data sets, denoted ML(MI).

			Complete	Incomplete		
Variable	Group	Week	ML(C)	ML(UN)	ML(IND)	ML(MI)
Group × Week	S	1	−11.406	−11.276	−11.372	−11.279
			(1.120)	(1.213)	(1.325)	(1.216)
Group × Week	S	4	−8.824	−8.985	−9.120	−8.986
			(1.153)	(1.189)	(1.342)	(1.210)
Group × Week	S	6	−3.152	−2.937	−4.669	−3.030
			(1.257)	(1.532)	(2.021)	(1.542)

relatively high. This trend is apparent in the sample means reported in Table 18.3, where the mean in the succimer group at week 6, based on the incomplete data, is approximately 1.5 units lower than the corresponding mean in the complete data set.

Next we compare and contrast a number of alternative methods for analyzing the incomplete data set on blood lead levels. First we consider a likelihood-based analysis of response profiles (see Sections 5.1–5.3) of the incomplete data. Recall that maximum likelihood (ML) estimation of the regression coefficients is valid when data are MAR provided that the assumed multivariate normal distribution for the responses has been correctly specified. This requires correct specification of not only the model for the mean response but also the model for the covariance among the responses. Because the analysis of response profiles gives unrestricted estimates of the means in each group, and also assumes an unstructured covariance among the responses, a likelihood-based analysis of the incomplete data should yield unbiased estimates of the regression coefficients. The results of an analysis of the incomplete data, summarized in terms of the three single-degree-of-freedom contrasts for the group × time interaction, are presented in Table 18.4; for ease of comparison the corresponding results for the complete data are also reproduced in Table 18.4. For simplicity we focus on the treatment group effect on changes in the mean blood lead levels from baseline to week 6. With almost 50% of the responses missing at week 6, this is the effect most likely to be sensitive to missingness. The analysis of response profiles of the incomplete data yields an estimate of −2.937 (with SE = 1.532) for the treatment group effect on changes in the mean blood lead levels from baseline to week 6. This is quite similar to the corresponding estimate from the complete data, albeit with a standard error that is discernibly larger. The larger standard error is to

be expected and reflects the loss of information due to the relatively large fraction of missing data at week 6. The estimates of the remaining two contrasts for the group $\times$ time interaction are also quite similar to those obtained from the complete data. In general, we would expect to obtain similar estimates of effects under the assumption that missingness is MAR and that the models for both the mean and the covariance have been correctly specified. In this case the former assumption is known to be valid by definition of the missing data mechanism that created the incomplete data set. The latter assumption seems tenable given that the analysis of response profiles makes few assumptions about the structure of the mean response and the covariance among the responses.

It is instructive to examine the sensitivity of ML estimation with incomplete data to misspecification of the likelihood. Specifically, we consider an analysis of response profiles under the naive assumption of constant variance and independence among the repeated responses, that is, under misspecification of the covariance. This analysis yields an estimate of -4.669 (with SE = 2.021) for the treatment group effect on changes in the mean blood lead levels from baseline to week 6; note, to correct for misspecification of the covariance, the standard errors reported in Table 18.4 are based on the empirical (or "sandwich") variance estimator. The estimate of -4.669 is approximately 60% larger (in absolute value) than the estimate obtained assuming an unstructured covariance; it is also approximately 50% larger than the estimate obtained from the complete data. This reflects the bias that can be introduced in a likelihood-based analysis of incomplete data when the model for the covariance is not correctly specified, even when missingness is truly MAR. As was noted earlier (see Section 18.1), when there are missing data, it is important to correctly specify the models for both the mean and the covariance to ensure that the regression parameters for the mean model are estimated without bias due to missingness.

A final approach to the analysis of the incomplete data is to use multiple imputation to replace any missing values by a set of m plausible values. Because the missingness patterns are non-monotone and the response is continuous, we use Markov chain Monte Carlo (MCMC) methods, based on a multivariate normality assumption, to impute missing values. Specifically, missing values are replaced by a set of $m = 50$ imputed values sampled from a chain that first uses 5000 "burn-in" iterations to achieve the desired stationary distribution. That is, 5000 burn-in iterations are executed before the first imputation from the chain is obtained. In addition, to remove any dependence between consecutive imputations, samples from the stationary distribution are drawn 50 times at every subsequent 500 iterations of the chain. Results of analyses of response profiles of the 50 completed data sets are then appropriately combined to yield the estimates and standard errors reported in the last column of Table 18.4.

The analysis based on multiple imputation yields an estimate of -3.030 (with SE = 1.542) for the treatment group effect on changes in the mean blood lead levels from baseline to week 6. This estimate and standard error are very similar to those yielded by the likelihood-based analysis of response profiles of the incomplete data. In general, we might expect the analysis based on multiple imputation to yield similar results to those from a likelihood-based analysis of the incomplete data because both sets of analyses are based on the assumption of multivariate normality.

Clinical Trial of Contracepting Women

Next we illustrate some of the methods described earlier for handling dropouts when the response is categorical rather than continuous. The methods are applied to data on a binary response from the longitudinal clinical trial of contracepting women (Machin et al., 1988) discussed in Sections 14.7 and 15.5. Recall that the goal of this trial was to compare the two treatments (100 mg or 150 mg of DMPA) in terms of how the rates of amenorrhea change over time with continued use of the contraceptive method. That is, the main interest is in an analysis that compares the rates of amenorrhea over time if those women who dropped out had remained on their assigned treatment. This is sometimes called an *explanatory* analysis (Schwartz and Lellouch, 1967). An "explanatory analysis," often referred to as an "as treated" analysis, focuses on what is thought to be the true underlying biological effects of the different treatments.

In this clinical trial a total of 1151 women completed menstrual diaries, and the diary data were used to generate a binary sequence for each woman, indicating whether or not she had experienced amenorrhea in four successive intervals. A feature of this trial is that there was substantial dropout. When the dropout rates are broken down by dosage group, the rates were marginally higher in the 150 mg dose group. Among women randomized to 100 mg (150 mg) of DMPA, 37% (39%) dropped out before the completion of the trial, 17% (17%) dropped out after receiving only one injection of DMPA, 12% (15%) dropped out after receiving only two injections, and 8% (6%) dropped out after receiving three injections. For women who dropped out before the end of the 3-month interval between injections, a determination of whether or not they experienced amenorrhea was made, on a proportionate basis, using their existing menstrual diary data for that interval.

Letting $Y_{ij} = 1$ if the i^{th} woman experienced amenorrhea in the j^{th} injection interval, we consider the following logistic regression model for the marginal probability of amenorrhea:

$$\text{logit}(\mu_{ij}) = \beta_1 + \beta_2 \text{t}_{ij} + \beta_3 \text{t}_{ij}^2 + \beta_4 \text{Dose}_i + \beta_5(\text{t}_{ij} \times \text{Dose}_i) + \beta_6(\text{t}_{ij}^2 \times \text{Dose}_i),$$

where $\mu_{ij} = \Pr(Y_{ij} = 1)$, $\text{t} = 0, 1, 2, 3$ for the four consecutive 3-month injection intervals, $\text{Dose} = 1$ if randomized to 150 mg of DMPA, and $\text{Dose} = 0$ otherwise. First we consider complete-case and available-data analyses of the data. If dropout is completely at random, then valid estimates of the marginal regression parameters can be obtained using a standard generalized estimating equations (GEE) approach. To account for the within-subject association among the repeated measures, we fit six separate pairwise log odds ratios. We note that the empirical and model-based standard errors are very similar in all of the analyses. For illustrative purposes only, we also consider a GEE analysis using LVCF imputation of the missing data. The differences in results are more easily discerned by considering the dose-specific estimated rates of amenorrhea in each of the injection intervals for the quadratic trend model given above (see Table 18.5).

Overall, the results of the complete-case and available-data GEE analyses suggest that the rates of amenorrhea in the second and third injection intervals are significantly higher for women who received the higher dose of DMPA, although these differences

Table 18.5 Estimated marginal rates of amenorrhea for quadratic trend model using GEE under three different methods for handling dropouts: complete-case (CC), available-data (AD), and last value carried forward (LVCF).

Method	Time	100 mg	150 mg	Difference	SE	Z
CC	3 months	0.176	0.155	−0.021	0.027	−0.79
	6 months	0.258	0.317	0.059	0.028	2.07
	9 months	0.369	0.463	0.094	0.033	2.83
	12 months	0.502	0.540	0.038	0.037	1.03
AD	3 months	0.184	0.201	0.017	0.023	0.73
	6 months	0.274	0.363	0.089	0.025	3.55
	9 months	0.388	0.499	0.111	0.030	3.68
	12 months	0.517	0.572	0.055	0.036	1.52
LVCF	3 months	0.184	0.201	0.017	0.023	0.75
	6 months	0.263	0.344	0.081	0.024	3.43
	9 months	0.350	0.453	0.103	0.027	3.78
	12 months	0.437	0.498	0.061	0.029	2.10

tend to decline by the end of the study. For example, during the third injection interval (6–9 months post-randomization) the predicted rates of amenorrhea from the available-data analysis are 0.499 in the 150 mg dose group and 0.388 in the 100 mg dose group. However, by the final follow-up visit there is no longer a discernible treatment difference, with predicted rates of amenorrhea of 0.572 in the 150 mg dose group and 0.517 in the 100 mg dose group.

The GEE analysis based on LVCF imputation produces discernibly lower estimated rates of amenorrhea during the third and fourth intervals, when compared to the available-data analysis, although the estimates of the treatment comparisons are not too dissimilar; however, the latter cannot be expected in general. Because LVCF uses a single imputation and does not reflect any uncertainty in the imputation, the standard errors for the estimated treatment comparisons are too small. Consequently, in contrast to the other methods, the analysis based on LVCF suggests that there are treatment differences in the estimated rates of amenorrhea at the end of the trial.

Note that if dropout is not completely at random, the complete-case and available-data GEE analyses of these data can yield biased estimates of the effects of treatment. Next we consider handling dropout using inverse probability weighted and multiple imputation methods. Recall that inverse probability weighted methods require a

Table 18.6 Estimated regression coefficients and standard errors from logistic regression model for the probability of remaining in the study.

Variable	Estimate	SE	Z
Intercept	1.668	0.104	15.98
I(t = 2)	0.137	0.119	1.15
I(t = 3)	0.729	0.144	5.06
Dose	0.068	0.131	0.52
Y_{ij-1}	−0.451	0.162	−2.79
Dose × Y_{ij-1}	−0.238	0.220	−1.08

model for the probability of dropout. We considered the following logistic regression model that assumes the log odds of remaining in the study (or, conversely, of dropout) depends on the previous observed response. Specifically, the model for being observed at the j^{th} occasion is given by

$$\text{logit}(\pi_{ij}) = \theta_1 + \theta_2\, I(t=2) + \theta_3\, I(t=3) + \theta_4\, \text{Dose}_i + \theta_5\, Y_{ij-1} + \theta_6\, (\text{Dose}_i \times Y_{ij-1}),$$

where $\pi_{ij} = \Pr(R_{ij} = 1 | R_{i1} = \cdots = R_{i,j-1} = 1, Y_{ij-1}, \text{Dose}_i)$. In this model the log odds of remaining in the study is allowed to vary over measurement occasions, to depend on dose group, and to depend on the previous observed response; the latter dependence is also allowed to vary by dose group.

By creating a "stacked" data set, estimates of the logistic regression parameters, $(\theta_1, ..., \theta_6)$, can be obtained using standard statistical software for fitting logistic regression. That is, we can create a "stacked" data set in which each individual contributes a sequence of binary "outcomes" to the logistic regression analysis. In this analysis each binary "outcome," R_{ij}, indicates whether the response was observed at a given occasion, from the second occasion (because $R_{i1} = 1$ for all individuals) until either the occasion when dropout occurred or the last intended measurement occasion. Thus study "completers" contribute a sequence of three binary responses ($R_{i2} = R_{i3} = R_{i4} = 1$) to the logistic regression analysis. Individuals who dropped out at the fourth occasion also contribute three binary responses ($R_{i2} = R_{i3} = 1, R_{i4} = 0$) to the analysis. In contrast, individuals who dropped out at the third occasion contribute only two binary responses ($R_{i2} = 1, R_{i3} = 0$), while individuals who dropped out at the second occasion contributes a single binary response ($R_{i2} = 0$).

The results of fitting the logistic regression model to this stacked data set are presented in Table 18.6. There is strong evidence that the probability of dropout is related to the previous response, although this dependence does not vary significantly between the dose groups. Specifically, for individuals in the 100 mg dose group, the conditional odds of dropout is approximately 60% higher (exp(0.451) = 1.57)

if they experienced amenorrhea at the previous occasion. Similarly, for those in the 150 mg dose group, the conditional odds of dropout is approximately two times higher (exp(0.451 + 0.238) = 1.99) if they experienced amenorrhea at the previous occasion. The estimated logistic regression coefficients can be used to obtain the estimated conditional probability of remaining in the study at each occasion for each individual, $\widehat{\pi}_{ij}$. The required weight w_{ij} for the i^{th} individual at the j^{th} occasion is then estimated by

$$\widehat{w}_{ij} = (\widehat{\pi}_{i1} \times \widehat{\pi}_{i2} \times \cdots \times \widehat{\pi}_{ij})^{-1}.$$

Because the first response was fully observed, with $R_{i1} = 1$ for all individuals, $\pi_{i1} = 1$ by definition. Thus, at the first occasion the weight is fixed at 1 for all individuals, while at all subsequent occasions the weights are given by $\widehat{w}_{ij}$ for occasions when the response is observed ($R_{ij} = 1$) and weights of zero for occasions when the response is not observed ($R_{ij} = 0$). Prior to conducting an IPW-GEE analysis, we examined the distribution of the estimated weights for the presence of discernibly large weights. The estimated weights ranged from 1.0 to 2.1, so there was no concern that a small subset of the observations might have undue influence on the analysis.

To conduct IPW-GEE estimation of the logistic regression model for the marginal probability of amenorrhea,

$$\text{logit}(\mu_{ij}) = \beta_1 + \beta_2 \text{t}_{ij} + \beta_3 \text{t}_{ij}^2 + \beta_4 \text{Dose}_i + \beta_5(\text{t}_{ij} \times \text{Dose}_i) + \beta_6(\text{t}_{ij}^2 \times \text{Dose}_i),$$

these occasion-specific weights need to be incorporated in the analysis. This can be achieved, for example, using the WEIGHT statement within PROC GENMOD in SAS or the `pweight` option for the `glm` command in Stata. However, to ensure that the weights are appropriately incorporated in the IPW-GEE analysis, it is necessary to make a "working independence" assumption for the within-subject association among the responses. Because a "working independence" assumption is made, standard errors are based on the "sandwich" variance estimator. The results of the IPW-GEE analysis are presented in Table 18.7. As in Table 18.6, the results are expressed in terms of the dose-specific estimated rates of amenorrhea in each of the injection intervals from the fitted quadratic trend model. The results from the IPW-GEE analysis are very similar to those obtained from the available-data analysis. Both the point estimates of the rates of amenorrhea and their standard errors are similar. The treatment group difference in the rates of amenorrhea at the fourth injection interval yielded by the IPW-GEE analysis is marginally lower than the corresponding difference obtained from the available-data analysis (4.9% versus 5.5%, respectively). Under the assumption that dropout is at random, and the model for dropout has been correctly specified, the results of the IPW-GEE analysis suggest that the rates of amenorrhea in the second and third injection intervals are significantly higher for those women who received the higher dose of DMPA, although these differences tend to decline by the end of the study.

Finally, we consider an analysis of the amenorrhea data that handles dropout using multiple imputation. Because the response is binary, imputation methods that are either explicitly or implicitly based on a multivariate normal distribution assumption for the response vector are unappealing. Instead, we can use a logistic regression

Table 18.7 Estimated marginal rates of amenorrhea for quadratic trend model using GEE under two different methods for handling dropouts: inverse probability weighted (IPW) GEE, and multiple imputation (MI) using logistic regression.

Method	Time	100 mg	150 mg	Difference	SE	Z
IPW	3 months	0.183	0.200	0.017	0.023	0.73
	6 months	0.276	0.361	0.084	0.026	3.29
	9 months	0.393	0.496	0.104	0.031	3.34
	12 months	0.521	0.570	0.049	0.037	1.33
MI	3 months	0.183	0.199	0.016	0.026	0.60
	6 months	0.276	0.365	0.089	0.026	3.43
	9 months	0.391	0.504	0.113	0.030	3.82
	12 months	0.518	0.579	0.061	0.033	1.85

imputation. Specifically, we consider a sequence of logistic regression models at the second through fourth occasion that assume the log odds of amenorrhea at each occasion to depend on all past observed responses, dose group, and their interactions. For example, the model at the third occasion is

$$\begin{aligned}\text{logit}\{\Pr(Y_{i3}=1|Y_{i1},Y_{i2},\text{Dose}_i)\} &= \theta_1+\theta_2\,\text{Dose}_i+\theta_3\,Y_{i1}+\theta_4\,Y_{i2}\\ &\quad+\theta_5\,\text{Dose}_i\times Y_{i1}+\theta_6\,\text{Dose}_i\times Y_{i2}.\end{aligned}$$

Three separate logistic regression models are fit to the data at the second through fourth occasion. Based on the estimated model parameters at each occasion, new logistic regression models are obtained by randomly drawing logistic regression parameters from their "posterior distribution." Finally, the missing binary responses at that occasion are then imputed from Bernoulli distributions with probabilities of amenorrhea determined by the randomly drawn logistic regression parameters. Logistic regression imputation of the remaining missing values continues in a similar manner until a completed data set has been created. To create 25 imputations, these steps are repeated 25 times, and results from the 25 analyses of the filled-in data sets are appropriately combined.

The results from the analysis based on multiple imputation (see bottom of Table 18.7) are similar to those obtained from the available-data analysis (see Table 18.5); they are also similar to those obtained from the IPW-GEE analysis. The treatment group difference in the rates of amenorrhea at the fourth injection interval yielded by the multiple imputation analysis is marginally higher than the corresponding difference obtained from the IPW-GEE analysis (6.1% versus 5.5%, respectively). This confluence of the estimates from the available-data, IPW-GEE, and multiple impu-

tation analyses provides some degree of reassurance that the main conclusions of the treatment group comparisons are not very sensitive to the methods used to handle dropout. The results from the multiple imputation analysis confirm the earlier findings that the rates of amenorrhea in the second and third injection intervals are significantly higher for those women who received the higher dose of DMPA, although these differences tend to decline by the end of the study.

18.5 "SANDWICH" VARIANCE ESTIMATOR ADJUSTING FOR ESTIMATION OF WEIGHTS*

In earlier sections we mentioned that the weights used in IPW-GEE are not known but must be estimated from the data at hand. It is possible to adjust the conventional standard errors for $\widehat{\beta}$, based on the "sandwich" variance estimator, to account for the estimation of the weights used by IPW-GEE. Indeed, such an adjustment will, in general, result in smaller standard errors. In this section we briefly outline how to construct the "sandwich" variance estimator for IPW-GEE that adjusts for estimation of the weights.†

Recall that the IPW-GEE estimator of β is obtained as the solution to the following weighted estimating equations:

$$S(\beta) = \sum_{i=1}^{N} S_i(\beta) = \sum_{i=1}^{N} D_i' V_i^{-1} W_i \left(Y_i - \mu_i\right) = 0.$$

If the estimation of the weights in W_i is completely ignored, and W_i is assumed fixed and known, then the usual formula for the "sandwich" variance estimator is

$$\mathrm{Cov}(\widehat{\beta}) = \left(\sum_{i=1}^{N} D_i' V_i^{-1} W_i D_i\right)^{-1} \left(\sum_{i=1}^{N} S_i(\beta) S_i'(\beta)\right) \left(\sum_{i=1}^{N} D_i' V_i^{-1} W_i D_i\right)^{-1}.$$

In practice, a logistic regression analysis (or any other suitable model) provides estimates of θ and the π_{ij}'s; thus the weights actually used in IPW-GEE are *estimated* rather than known. Specifically, the estimates of θ (and the π_{ij}'s), and hence the weights w_{ij}, are obtained as the solution to the following estimating equations:

$$S(\theta) = \sum_{i=1}^{N} S_i(\theta) = \sum_{i=1}^{N} \sum_{j=2}^{n} R_{i,j-1} Z_{ij} (R_{ij} - \pi_{ij}) = 0.$$

These are the equations for the logistic regression analysis of the "stacked" data set. The solution to these estimating equations provides estimates of θ and the π_{ij}'s; the

†This section provides the formula for adjusting the "sandwich" variance estimator for estimation of the weights in IPW-GEE. The content of this section is somewhat technical and can be omitted without loss of continuity.

required weights w_{ij} are then estimated as

$$\widehat{w}_{ij} = (\widehat{\pi}_{i1} \times \widehat{\pi}_{i2} \times \cdots \times \widehat{\pi}_{ij})^{-1}.$$

The IPW-GEE estimate of β is then obtained using *estimated* weights, denoted by $W_i(\widehat{\theta})$. In large samples the resulting IPW-GEE estimator has a multivariate normal distribution, with mean β, and estimator of its covariance given by the *adjusted* "sandwich" variance estimator (where the adjustment is for estimation of the weights),

$$\text{Cov}(\widehat{\beta}) = \left(\sum_{i=1}^{N} D_i' V_i^{-1} W_i(\widehat{\theta}) D_i\right)^{-1} \left(\sum_{i=1}^{N} S_i^*(\beta,\theta) S_i^{*\prime}(\beta,\theta)\right) \left(\sum_{i=1}^{N} D_i' V_i^{-1} W_i(\widehat{\theta}) D_i\right)^{-1},$$

where

$$S_i^*(\beta,\theta) = S_i(\beta) - \left(\sum_{i=1}^{N} S_i(\beta) S_i'(\theta)\right) \left(\sum_{i=1}^{N} S_i(\theta) S_i'(\theta)\right)^{-1} S_i(\theta).$$

Note that $S_i^*(\beta,\theta)$ is the residual from a multivariate linear regression of $S_i(\beta)$ on $S_i(\theta)$. The replacement of $S_i(\beta)$ by $S_i^*(\beta,\theta)$ in the "meat" (or center-piece) of the "sandwich" variance estimator yields an appropriate adjustment to the standard errors of $\widehat{\beta}$ for estimation of the weights in the IPW-GEE method. Moreover, because $S_i^*(\beta,\theta)$ is a residual, by definition, the "residual sums of squares" in the "meat" of this "sandwich" variance estimator must be smaller than (or equal to) the corresponding "sums of squares" based on $S_i(\beta)$ in the conventional "sandwich" variance estimator. This "sums of squares" argument helps explain why the *adjusted* "sandwich" variance estimator yields smaller standard errors for $\widehat{\beta}$.

18.6 COMPUTING: MULTIPLE IMPUTATION USING PROC MI IN SAS

As described in earlier sections of this chapter, making inferences from a statistical analysis based on multiple imputation is a three-step process. First, the missing data are filled-in m times to create m completed data sets. Second, the m completed data sets are analyzed using standard statistical methods (e.g., fitting linear models using PROC MIXED in SAS). Finally, the results from the m analyses of the completed data sets are appropriately combined. In this section we focus on the first step and briefly discuss how to create m completed data sets using PROC MI in SAS. There is a complementary procedure in SAS, PROC MIANALYZE, for combining the results from the m analyses of the completed data sets.

The MI procedure in SAS is a general procedure for multiple imputation that offers a number of alternative methods for imputing missing data, especially when the patterns of missing data are monotone. No attempt is made here to give a comprehensive review of the main features of PROC MI. Instead, we present illustrative source code for generating m completed data sets using a number of alternative methods for imputation.

Before discussing the command syntax for PROC MI, we note that when imputing longitudinal data with missing responses, it is important to structure the data set appropriately. To capitalize on the correlation among the repeated measurements of the responses, the procedure requires that each repeated measurement in a longitudinal data set be a separate variable rather than a separate "record." That is, for the purposes of imputing missing values, PROC MI should be applied to a data set that is structured in a "wide" rather than a "long" format, with a single "record" for each individual. In a "wide" format, the imputation model for a missing response at any particular occasion can include as predictors the responses at any of the remaining occasions, thereby capitalizing on the positive correlation among the repeated measurements. After m completed data sets have been generated, each of these "wide" format data sets can then be transformed to a "long" format data set prior to analysis using standard statistical methods (e.g., PROC MIXED or PROC GENMOD in SAS). Finally, we note that when using imputation methods that rely on a multivariate normal assumption, imputations can often be improved by transforming the response variable. For example, when the longitudinal responses are quantitative but have distributions that are not symmetric, transformation of the response (e.g., log transformation of variables with positive skewed distributions) should improve the imputation. After the values have been imputed on the transformed scale, these can be reverse-transformed back to the original scale (e.g., if a response variable has been log transformed prior to imputation, the imputed values can subsequently be exponentiated). Of course, the multivariate normal assumption has no consequence for those variables that have no missing data but are included in the imputation process (e.g., treatment or exposure group indicators).

To use multiple imputation to generate 25 completed data sets in the setting where there is intermittent (non-monotone) missingness and the vector of response is assumed to have a multivariate normal distribution, we can use the illustrative SAS commands given in Table 18.8. By default, PROC MI in SAS uses MCMC methods for generating imputations when missingness is non-monotone. In contrast, when the missing data patterns are monotone there are three alternative methods for imputation: (1) regression method, (2) predictive mean matching, and (3) propensity score method. With binary (or ordinal) responses, the illustrative SAS commands in Table 18.9 can be used to create 25 completed data sets using logistic regression imputations. Below we present a brief description of the command statements used in Tables 18.8 and 18.9. For a more detailed description of these command statements, and other statements and options, the reader is referred to the extensive SAS documentation for PROC MI.

PROC MI <options>;

The PROC MI statement calls the procedure MI in SAS. It include options for the input SAS data-set to be read in (DATA=*SAS-data-set*) and for the creation of an output SAS data-set (OUT=*SAS-data-set*) that contains the completed data sets with imputed values. The output SAS data-set includes an additional index variable, _IMPUTATION_, to identify the imputation number. For each imputation, the output data set contains all the variables in the input data set, with miss-

Table 18.8 Illustrative commands for multiple imputation via MCMC method using PROC MI in SAS.

```
PROC MI DATA=widefile SEED=364865 NIMPUTE=25 OUT=mifile;
    VAR  group  y1  y2  y3  y4;
    MCMC NBITER=5000 NITER=500;
```

Table 18.9 Illustrative commands for multiple imputation via logistic regression using PROC MI in SAS.

```
PROC MI DATA=widefile SEED=364865 NIMPUTE=25 OUT=mifile;
    VAR  group  y1  y2  y3  y4;
    CLASS  group  y1  y2  y3  y4;
    MONOTONE  LOGISTIC(y2=group y1 group*y1);
    MONOTONE  LOGISTIC(y3=group y1 y2 group*y1 group*y2);
    MONOTONE  LOGISTIC(y4=group y1 y2 y3 group*y1 group*y2 group*y3);
```

ing values replaced by imputed values. The PROC MI statement also includes options for the number of imputations to be created (NIMPUTE=*number*) and a seed used to initialize the random number generator (SEED=*number*); the latter is useful to ensure that results can be duplicated in a later run.

VAR variables;

The VAR statement lists the variables to be used in the imputation process. When the patterns of missingness are monotone, the order of variables in the VAR statement is important; they should be listed in order of the variables that are fully observed, followed by the variable with the fewest missing values, then the variable with the second fewest missing values, and so on. Note that the VAR statement can include variables that are fully observed but thought to be predictive of the probability of missingness and/or predictive of the missing responses.

CLASS variables;

The CLASS statement is used to define categorical variables in the VAR statement; these categorical variable can be used either as predictors for imputed variables or as imputed variables for data sets with monotone missing patterns. The CLASS statement must be used in conjunction with the MONOTONE statement (described later).

MCMC <options>;

The MCMC statement uses a Markov chain Monte Carlo method to impute values for a data set with an arbitrary missing pattern, under an assumed multivariate normal distribution for the variables. It is also the default method of imputation when neither the MCMC nor the MONOTONE statement is specified. It includes options for controlling the number of burn-in iterations before the first imputation in each chain (NBITER=*number*) and the number of iterations between imputations in a chain (NITER=*number*).

MONOTONE <options>;

The MONOTONE statement specifies that the missingness patterns are monotone and provides three methods for imputing a quantitative variable: (1) regression method (REGRESSION), (2) predictive mean matching (REGPMM), and (3) propensity score method (PROPENSITY). These three options for the MONOTONE statement are:

REGRESSION <*imputed* = *effects*>
REGPMM <*imputed* = *effects*>
PROPENSITY <*imputed* = *effects*>

With the MONOTONE statement, the variables are imputed sequentially in the order given by the VAR statement. The effects specification option allows you to use a different set of predictors for each imputed variable (see Table 18.9). When the effect specification option is not used, the preceding variables in the VAR statement are used as the default predictors when imputing missing values for any particular variable. You can also specify more than one method of imputation in the MONOTONE statement, and for each imputed variable, the predictors can be specified separately using the effects specification option.

The MONOTONE statement provides the following options for imputing a categorical variable (defined on the CLASS statement): (1) logistic regression (LOGISTIC) imputation of a binary or ordinal variable and (2) discriminant analysis (DISCRIM) of a nominal categorical variable. The effects specification option can also be used with both of these methods of imputation. The discriminant analysis imputation is based on assuming that within levels of the categorical variable the predictors have a multivariate normal distribution with means that vary across the categories (but with a constant covariance matrix); therefore it requires that all predictors of the missing categorical variable be continuous. In contrast, the predictors in a logistic regression imputation can be a mixture of continuous and categorical variables.

Finally, once the missing data are filled in m times, the m completed data sets can be analyzed using standard statistical methods. For example, the 25 completed data sets created in Table 18.8 can be analyzed by fitting linear models using PROC MIXED in SAS. The results from these multiple analyses of the completed data sets can then be appropriately combined. Table 18.10 presents illustrative commands for combining the results of analyses of multiple imputed data sets using PROC MIANALYZE in SAS. The first set of SAS commands in Table 18.10 are used to convert the imputed

Table 18.10 Illustrative commands for combining the results of analyses of multiple imputed data using PROC MIANALYZE in SAS.

```
DATA milong;
    SET mifile;
    y=y1; time=0; OUTPUT;
    y=y2; time=1; OUTPUT;
    y=y3; time=2; OUTPUT;
    y=y4; time=3; OUTPUT;

PROC SORT;
    BY _IMPUTATION_;

PROC MIXED DATA=milong;
    CLASS id;
    MODEL y = group time group*time / S COVB;
    RANDOM INTERCEPT time / SUBJECT=id TYPE=UN;
    BY _IMPUTATION_;
    ODS OUTPUT SOLUTIONF=beta COVB=varbeta;

PROC MIANALYZE PARMS=beta;
    MODELEFFECTS INTERCEPT group time group*time;
```

data set (here named `mifile`) to a "long format" data set (`milong`) required for longitudinal analyses. Prior to analysis, the "long format" data set should be sorted by the index variable `_IMPUTATION_`. The next set of SAS commands use PROC MIXED to fit a linear mixed effects model, with randomly varying intercepts and slopes, to each of the imputed data sets. The BY statement in PROC MIXED is used to generate separate analyses based on observations grouped by the index variable `_IMPUTATION_`; this produces a separate analysis for each imputed data set. The ODS OUTPUT statement is used to create SAS data-sets containing the regression parameter estimates (`beta`) and the covariance matrices of the regression parameter estimates (`varbeta`) from the analyses of the imputed data sets. The final set of SAS commands use PROC MIANALYZE to appropriately combine these estimates to yield a single estimate of the regression parameters, together with standard errors that reflect the uncertainty inherent in the imputation of the missing data. The PARMS statement in PROC MIANALYZE is used to name the SAS data-set containing the regression parameter estimates (and the associated standard errors) from the analyses of the

Table 18.11 Data for three subjects from the Clinical Trial of Contracepting Women.

id	dose	time	y	prevy	r
21	0	0	0	.	1
21	0	1	.	0	0
329	1	0	1	.	1
329	1	1	0	1	1
329	1	2	.	0	0
962	0	0	0	.	1
962	0	1	1	0	1
962	0	2	1	1	1
962	0	3	1	1	1

imputed data sets (here named `beta`); for multivariate inferences (e.g., multivariate Wald tests), the SAS data-set containing the covariance matrices of the regression parameter estimates (`varbeta`) must also be provided. For a more detailed description and explanation of command statements, and other options, the reader is referred to the extensive SAS documentation for PROC MIANALYZE.

18.7 COMPUTING: INVERSE PROBABILITY WEIGHTED (IPW) METHODS IN SAS

In principle, the application of inverse probability weighted (IPW) methods is straightforward because most statistical procedures allow for the inclusion of sampling weights. Specifically, the method can be implemented as a two-step process. The first step requires the calculation of the inverse probability weights by fitting a suitable model (e.g., logistic regression) for the probability of remaining in the study at each occasion (or, conversely, the probability of dropout). The second step requires the fitting of standard procedures for longitudinal analyses, where each observation is weighted by the inverse probabilities estimated in the first step. In the following we provide sample SAS commands for estimating the inverse probability weights and for replicating the weighted GEE analysis of the amenorrhea data from the *Clinical Trial of Contracepting Women* described in Section 18.4. The data for three subjects from the *Clinical Trial of Contracepting Women* are displayed in Table 18.11.

The first set of SAS commands in Table 18.12 are used to read a "long format" data set, with separate records for each measurement occasion. The data set contains a variable, `r`, with `r` $= 1$ if the outcome (`y`) is observed, and `r` $= 0$ if missing due

Table 18.12 Illustrative commands for calculating cumulative probabilities of remaining in the study based on a logistic regression using PROC GENMOD in SAS.

```
DATA contracep;
    INFILE 'contracep.dat';
    INPUT id dose time y prevy r;

PROC SORT DATA=contracep;
    BY id time;

PROC GENMOD DESCENDING;
    CLASS time (PARAM=REF REF="1");
    MODEL r = time dose prevy dose*prevy / DIST=BIN;
    WHERE time NE 0;
    OUTPUT OUT=predict P=probs;

PROC SORT DATA=predict;
    BY id time;

DATA wgt (KEEP=id time cumprobs probs);
    SET predict;
    BY id time;
    RETAIN cumprobs;
    IF FIRST.id then cumprobs=probs;
    ELSE cumprobs=cumprobs*probs;
```

to dropout. In this illustration there are no missing data at baseline (`time` $= 0$). The data set also contains a variable, `prevy`, denoting the value of the outcome at the previous occasion; by definition, `prevy` is missing at baseline (`time` $= 0$) but is observed at all post-baseline occasions up to and including the time of dropout. The latter variable is used in the model for the inverse probability weights.

The next set of SAS commands in Table 18.12 use PROC GENMOD to fit a logistic regression model to the probability of remaining in the study at each occasion. The logistic regression is based on a "stacked" data set in which each individual contributes a sequence of binary "outcomes" (`r`) to the analysis. Each binary "outcome", denoted by R_{ij} in earlier sections of this chapter, is an indicator of whether the response was observed at a given occasion, from the second occasion (because, in this illustration, there are no missing data at baseline and $R_{i1} = 1$ for all subjects) until

Table 18.13 Illustrative commands for calculating inverse probability weights and for fitting a marginal logistic regression model via IPW-GEE using PROC GENMOD in SAS.

```
DATA combine;
    MERGE contracep wgt;
    BY id time;
    IF (time=0) THEN ipw=1;
    ELSE ipw=1/cumprobs;

PROC GENMOD DESCENDING DATA=combine;
    WEIGHT ipw;
    CLASS id;
    MODEL y = dose time time*time dose*time dose*time*time / DIST=BIN;
    REPEATED SUBJECT=ID / TYPE=IND;
```

either the occasion when dropout occurs or the last intended measurement occasion. The WHERE statement in PROC GENMOD restricts the analysis to the post-baseline occasions; the DESCENDING option ensures that the logistic regression models the probability that $\texttt{r} = 1$. The OUTPUT statement creates a new SAS data set named `predict` containing the estimated probabilities (`probs`) at each occasion. It is important that both the original SAS data-set (here named `contracep`) and the SAS data-set with the estimated probabilities (here named `predict`) are sorted by subject identification number (`id`), and within subject `id`, by measurement occasion (`time`).

The final set of SAS command in Table 18.12 are used to calculate the *cumulative* probabilities of remaining in the study at each occasion. This requires relatively sophisticated use of a DATA step, with both BY and RETAIN statements. In general, the use of the BY statement in a DATA step results in the creation of two temporary (so-called automatic) variables: FIRST.*variable* and LAST.*variable* (for any *variable* listed on the BY statement). In Table 18.12, `FIRST.id` takes the value 1 for the first observation within `id` and 0 for all other observations within `id` (similarly `LAST.id` takes the value 1 for the last observation within `id` and 0 for all other observations within `id`). The RETAIN statement is used here to "remember" values of a variable from a previous observation. This use of the RETAIN statement allows for simple calculation of the cumulative probabilities (`cumprobs`) at each occasion for every subject.

Finally, the set of SAS commands in Table 18.13 are used to: (1) merge the SAS data-set (`wgt`) containing the cumulative probabilities with the original SAS data-set (`contracep`), (2) create inverse probability weights, `ipw` (and set $\texttt{ipw} = 1$ at baseline becaue there were no missing data at baseline), and (3) use PROC GENMOD to fit a marginal logistic regression model using IPW-GEE. The WEIGHT statement

in PROC GENMOD weights each observation by the estimated inverse probability weights, `ipw`. To ensure that the weights are correctly applied within PROC GENMOD, a "working independence" assumption (TYPE=IND) for the within-subject association among the responses must be used. Because a "working independence" assumption is adopted, standard errors are based on the "sandwich" variance estimator.

18.8 FURTHER READING

General reviews of multiple imputation can be found in Rubin (1996), Rubin and Schenker (1991), Schafer (1999), and Horton and Lipsitz (2001). A comprehensive overview of the use of multiple imputation for handling incomplete data in longitudinal studies can be found in Chapters 9 and 13 of Molenberghs and Kenward (2007) and Chapter 28 of Molenberghs and Verbeke (2005). A detailed description of imputation by "chained equations" can be found in Raghunathan et al. (2001) and van Buuren (2007); also see van Buuren et al. (2006) for a study of the robustness of the method.

Inverse probability weighted methods have their roots in the survey sampling literature; IPW methods for longitudinal models were introduced in a landmark paper in the statistical literature by Robins et al. (1995). The article by Preisser, Lohman and Rathouz (2002) provides a concise but very useful summary of this topic in the context of dropout in longitudinal studies; also see Chapters 10 and 13 of Molenberghs and Kenward (2007).

Finally, missing data can arise due to death. Loss to follow-up due to death is qualitatively distinct from dropout due to other reasons and, ordinarily, needs to be handled quite differently in the analysis of longitudinal data; see Dufouil et al. (2004) for a very useful discussion of this topic.

Bibliographic Notes

Multiple imputation was introduced by Rubin (1978) as a general method for handling missing data; see Rubin (1987) for a book-length treatment of this topic. Inverse probability weighted methods were first proposed in the sample survey literature by Horvitz and Thompson (1952). Robins et al. (1995) developed an inverse probability weighted estimating (IPW) equations approach for handling missing data in longitudinal studies. A more detailed description of the theory underlying the IPW methodology can be found in the text by Tsiatis (2006). The connection between imputation and inverse probability weighted methods is discussed in Reilly and Pepe (1997). Finally, Javaras and Fitzmaurice (2009) discuss analytic methods for handling extraneous covariates that are potentially predictive of missingness and also related to the covariates of interest and the outcomes.