CHAPTER 18

Selection and pattern-mixture models

Roderick Little

Contents

18.1 Introduction: Theoretical framework .................................................. 409
18.2 Ignorable maximum likelihood methods .............................................. 413
18.3 Non-ignorable models for the joint distribution of $Y$ and $R$ ................. 416
18.4 Bivariate data with dropouts............................................................. 417
18.5 Mixed models with dropouts .............................................................. 422
  18.5.1 Covariate-dependent dropout .................................................. 422
  18.5.2 MAR dropout ................................................................. 423
  18.5.3 Non-ignorable outcome-dependent dropout ............................... 423
  18.5.4 Non-ignorable random-coefficient dependent dropout ................. 424
  18.5.5 Shared-parameter models ..................................................... 424
18.6 Mixed-effect pattern-mixture models ................................................. 426
18.7 Conclusion .............................................................................. 428
References ...................................................................................... 429

18.1 Introduction: Theoretical framework

Missing data are a common problem in longitudinal data sets, as the overview in Chapter 17 discussed. This chapter considers likelihood-based methods for handling this problem, based on parametric models for the data and missing-data mechanism. These models can also form the basis for multiple imputation approaches discussed in Chapter 21. Approaches based on estimating equations other than the likelihood, including inverse probability weighting methods, are discussed in Chapter 20. A useful tutorial that discusses both likelihood-based and estimating equations approaches is Hogan, Roy, and Korkontzelou (2004).

Unless missing data are a deliberate feature of the study design, it is important to try to limit them during data collection, since any method for compensating for missing data requires unverifiable assumptions that may or may not be justified. Since data are still likely to be missing despite these efforts, it is important to try to collect covariates that are predictive of the missing values, so that an adequate adjustment can be made. In addition, the process that leads to missing values should be determined during the collection of data if possible, since this information helps to model the missing-data mechanism when the incomplete data are analyzed.

We first briefly review parametric likelihood methods in the absence of missing data, as discussed in earlier chapters in this book. We suppose there are $N$ individuals, and $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$ is a vector of repeated measurements planned for individual $i$, and write $Y = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_N\}$. Also associated with individual $i$ at time $j$ is a $(p \times 1)$ vector of covariates $\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ijp})'$, with $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i})$ the resulting $(n_i \times p)$ matrix of covariates, and $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$. Likelihood-based methods assume a model for the distribution $f(Y|X, \gamma)$ of $Y$ given $X$ with unknown parameters $\gamma$. Assuming the
individuals $i$ are independent, this distribution factors into a product of distributions over the individuals $i$:

$$f(Y|X, \gamma) = \prod_{i=1}^{N} f(Y_i|X_i, \gamma),$$

where $f(Y_i|X_i, \gamma)$ is the distribution of $Y_i$ given $X_i$ (density function for continuous $Y_i$). The likelihood of $\gamma$ given data $\{(Y_i, X_i) : i = 1, \ldots, N\}$ is

$$L(\gamma|Y, X) = c \prod_{i=1}^{N} f(Y_i|X_i, \gamma),$$

considered as a function of the parameters $\gamma$, where $c$ is an arbitrary factor that does not depend on $\gamma$. The maximum likelihood (ML) estimate $\hat{\gamma}$ of $\gamma$ is the value that maximizes $L(\gamma|Y, X)$. Large-sample ML inferences under the model are based on the normal approximation

$$(\gamma - \hat{\gamma}) \sim N(0, C), \quad (18.1)$$

where $N(0, C)$ denotes the multivariate normal distribution with mean 0 and covariance matrix $C$, and $C$ is one of several estimates, for example the sample covariance matrix from the bootstrap distribution of the estimates, or the inverse observed information matrix $\{-\partial^2 \log L(\gamma|Y, X)/\partial \gamma \partial \gamma\}^{-1}$. Following Little and Rubin (2002), Equation (18.1) is written to have both a frequentist interpretation, where $\gamma$ is fixed, $\hat{\gamma}$ is random, and the equation represents the asymptotic sampling distribution of $\hat{\gamma}$, or a Bayesian interpretation, where $\hat{\gamma}$ is fixed, $\gamma$ is random, and the equation represents a large-sample approximation to the posterior distribution of $\gamma$. Bayesian inference adds a prior distribution $p(\gamma)$ for the parameters, and bases inference on the posterior distribution $p(\gamma|Y, X) = c p(\gamma) L(\gamma|Y, X)$, where $c$ is a normalizing constant.

Now suppose there are gaps in the data $Y$ on the repeated measures. I consider an unobserved value to be missing if there is a true underlying value that is meaningful for analysis. This may seem obvious, but is not always the case. For example, in a study of a behavioral intervention for people with heart disease, it is not meaningful to consider a quality of life measure to be missing for subjects who die prematurely during the course of the study. Rather, it is preferable to restrict the analysis to the quality of life measures of individuals while they are alive. This issue — whether values are truly considered missing or not — has implications for the choice of missing-data model, as discussed further below.

Let $R_{ij}$ be the missing-data indicator for $Y_{ij}$, with value 1 if $Y_{ij}$ is observed and 0 if $Y_{ij}$ is missing, and $R_i = (R_{i1}, \ldots, R_{in_i})$. The vector $Y^n_i$ denotes the set of observed values for individual $i$, and $Y^m_i$ the set of missing values. Unless stated otherwise, we assume that $X_i$ is observed for all $i$, so the covariates do not contain missing values. This is not an innocuous assumption; for example, if covariates are measured repeatedly over time, then they are typically also missing after an individual drops out. Some comments on the case where covariates are also missing are provided in Section 18.7. As in Chapter 17, the problem is then to make inferences about $\gamma$ based on the set of incomplete data $(R, Y^n, X) = \{(R_i, Y^n_i, X_i) : i = 1, \ldots, N\}$, rather than the complete data $(Y, X)$.

With incomplete data, in general we need a model for the joint distribution of $Y$ and $R$, with density $f(R, Y|X, \theta)$ indexed by parameters $\theta = (\gamma, \phi)$, where $\gamma$ characterizes the model for the data $Y$ and $\phi$ the model for the missing-data indicators $R$. Again assuming independence over individuals, this density can be written as $f(R, Y|X, \theta) = \prod_{i=1}^{N} f(R_i, Y_i|X_i, \theta)$. With no missing values, likelihood inferences would be based on the complete-data likelihood

$$L(\theta|R, Y, X) = c \prod_{i=1}^{N} f(R_i, Y_i|X_i, \theta),$$
INTRODUCTION

where, as before, \( c \) is an arbitrary constant independent of \( \theta \). With missing data, likelihood inferences are based on the observed-data likelihood, or simply the likelihood, given data \( \{(R_i, Y_i^o, X_i) : i = 1, \ldots, N\} \), which is obtained formally by integrating the missing data \( Y_i^m \) out of the density of \( (R_i, Y_i) \):

\[
L(\theta|R,Y^o,X) = c \prod_{i=1}^N \int f(R_i,Y_i|X_i,\theta)dY_i^m.
\]

(18.2)

In principle, inferences for \( \theta \) can then proceed in the same way as for inferences about \( \gamma \) in the case of complete data. That is, the ML estimate \( \hat{\theta} \) of \( \theta \) is the value that maximizes \( L(\theta|R,Y^o,X) \). Large-sample ML inferences under the model are based on the normal approximation

\[
(\theta - \hat{\theta}) \sim N(0, C_0),
\]

where \( C_0 \) is an estimate of the large-sample covariance matrix, for example the inverse observed information matrix \( -\partial^2 \log L(\theta|R,Y^o,X)/\partial \theta \partial \theta' \)^{-1}. Bayesian inference is based on the posterior distribution of \( \theta \), obtained by multiplying the likelihood \( L(\theta|R,Y^o,X) \) by a prior distribution \( p(\theta) \) for the parameters. ML estimates under a correctly specified model are fully efficient, and in particular make use of information in the incomplete cases that is lost when the incomplete cases are dropped from the analysis. The Bayesian approach shares the optimal large-sample properties of ML, and can yield better small-sample inferences. See, for example, Little and Rubin (2002, Chapter 6).

Missing data complicate likelihood-based inferences in a number of ways:

(i) As described above, in general a model for the joint distribution of \( R \) and \( Y \) is needed, rather than simply a model for the distribution of \( Y \). Specifying a model for \( R \) requires knowledge about the process leading to missing values, about which little is often known. Results for the parameters of interest \( \gamma \) tend to be sensitive to the assumptions contained in this model, so a bad specification of this model can lead to poor inferences, even if the model for \( Y \) is correctly specified. When the missing-data mechanism is ignorable for likelihood inference, inferences can be based on the ignorable likelihood

\[
L_{\text{ign}}(\gamma|Y^o,X) = c \prod_{i=1}^N \int f(Y_i|X_i,\gamma)dY_i^m = c \prod_{i=1}^N f(Y_i^o|X_i,\gamma).
\]

(18.3)

This function is generally much easier to deal with than the full likelihood (18.2). The integral in the latter complicates the computation; the ignorable likelihood (18.3) does not require a model for \( R \), which can be difficult to specify; furthermore, the parameters in the full likelihood tend to be at best weakly identified, making inference problematic.

For these reasons, most likelihood analyses for incomplete longitudinal data with dropouts or intermittent missingness are currently based on (18.3) rather than the full likelihood (18.2). As discussed in Chapter 17 and Chapter 20, the key condition for this simplification to occur is that the missing data are missing at random (MAR), in that missingness only depends on the data through observed values \( (Y^o,X) \):

\[
f(R_i|Y_i^o, X_i, \phi) = f(R_i|Y_i^o, X_i, \phi) \text{ for all } Y_i^m
\]

(18.4)

(Rubin, 1976; Little and Rubin, 2002, Chapter 6). There is also a more technical separability condition, which states that the parameters \( \gamma \) and \( \phi \) have distinct parameter spaces, but that condition is less important and often reasonable in applications. The next section considers ML methods based on (18.3) under the assumption that the missing-data mechanism is ignorable. I then consider models that incorporate mechanisms that are not missing at random (NMAR).
(ii) The key role of (18.4) in justifying an inference based on (18.3) makes MAR very desirable, and it is worth collecting covariate information $W_i$ that makes the MAR assumption plausible, and incorporating this information in the analysis. When it is appropriate to condition on $W_i$ in the final analysis, the likelihood-based analysis is straightforward, because $W_i$ is just incorporated in the covariate matrix $X_i$. When it is not appropriate to condition on $W_i$ in the final analysis, the likelihood-based methods considered here are trickier, because the $W_i$ then need to be assigned a distribution and integrated out for final inferences. While this is quite possible in principle, the multiple imputation methods in Chapter 21 provide a more convenient solution, because the imputations can be based on an imputation model that conditions on $W_i$, but the final analysis simply omits these variables.

(iii) The missing data may render some parameters in the likelihood function unidentified, in the sense that unique ML estimates are not available, or weakly identified. As a trivial example, the mean of $Y_{ij}$ is not identified if all individuals drop out before measure $j$. As a more complex example, Example 1 below concerns a longitudinal study measuring the growth of children between ages 3 and 12, which initially recruits children of various ages and follows them for at most 8 years, or until they reach age 12. The correlation between growth measures at ages 3 and 12 is not identified, since these measures are never recorded for the same child. A general covariance matrix for the repeated measures is thus not identified from this data structure.

Lack of identifiability yields inferential problems, such as estimates with a high degree of uncertainty, and computational problems, such as iterative ML algorithms failing to converge or converging painfully slowly. Such problems are particularly prevalent when the missing-data mechanism is non-ignorable. When parameters are poorly identified, it may be better to conduct a sensitivity analysis, where answers are computed for a range of plausible values of these parameters, rather than trying to estimate them from the data. We provide some examples of analyses of this kind below.

(iv) The missing data may increase the sensitivity of inferences to misspecification of the model for the data $Y$. For example, incorrectly assuming a linear relationship between an outcome and a covariate may lead to more serious bias when missingness depends on the value of the covariate than when it does not. Interestingly, the cases where inferences are sensitive to the model tend to be the cases where including the incomplete cases in the analysis has the greatest payoff, in terms of reduced bias and increased precision.

(v) The observed-data likelihoods are typically more complicated than likelihoods based on the complete data, and have greater potential to be multimodal and non-normal in shape. Thus, larger samples are needed for asymptotic methods like ML and information matrix based standard errors to be satisfactory, than is the case with complete data.

(vi) Missing data complicate the estimation of standard errors of parameters. In particular, the popular EM algorithm for finding ML estimates does not yield standard errors of parameters as part of its output. The information matrix of the parameters is more complicated than with complete data, and is harder to calculate and invert. For example, with many repeated-measures models, the complete-data information matrix is block-diagonal between parameters characterizing the mean and covariance structures. This means that if standard errors are only needed for parameters characterizing the mean structure, these can be obtained by inverting the submatrix of the information matrix corresponding to those parameters. With missing data, the block diagonal structure is lost under MAR, so the full matrix needs to be inverted. Alternative approaches to computing standard errors are to base standard errors on draws
from the Bayesian posterior distribution of estimates, or to compute bootstrap or jackknife standard errors (Little and Rubin 2002, Chapter 5 and Chapter 9). Bootstrap samples can be obtained as simple random samples with replacement of size $N$ from the sampled individuals, ignoring pattern. Stratifying on pattern is also a possibility, although gains from doing this seem unclear.

(vii) Computation of ML estimates and associated standard errors, and posterior distributions for Bayesian inference, is typically more challenging, requiring iterative methods such as the EM algorithm and Markov chain Monte Carlo simulation. I focus here on features of the models themselves rather than on computational aspects. For discussions of computation, see Tanner (1991), Schafer (1997), and Little and Rubin (2002). For many users not interested in developing their own programs, choices are often limited to missing-data methods that are available in widely available statistical software packages. These choices are increasing, but gaps still remain, as indicated below.

18.2 Ignorable maximum likelihood methods

ML inference based on the ignorable likelihood (18.3) is formally similar to ML with complete data: a model is not required for the missing-data mechanism, and in large samples, hypothesis tests and confidence intervals can be based on the ML estimates of parameters and asymptotic standard errors, as for complete data. We discuss the ignorable likelihood approach for a simple bivariate normal model with dropouts, a more general mixed model suitable for many repeated-measures problems, and extensions to non-normal models available in current statistical software.

**Model 1. A normal model for two repeated measures with MAR dropout.** Suppose there are just two repeated measures ($n_i = 2$ for all $i$), and $(Y_{i1}, Y_{i2})$ are observed and $R_i = 1$ for $i = 1, \ldots, r$, and $Y_{i1}$ is observed, $Y_{i2}$ is missing, and $R_i = 0$ for $i = r + 1, \ldots, N$. Let $\mu = (\mu_1, \mu_2)$, $\mu_j = E(Y_{ij})$, and $\Sigma = (\sigma_{jk})$ be the covariance matrix of $(Y_{i1}, Y_{i2})$: often the covariance matrix for repeated-measures data is assigned a special structure, such as compound symmetry, but here I assume this matrix is unrestricted. Suppose interest concerns the difference in means between the two time points, $\mu_{\text{diff}} = \mu_2 - \mu_1$. Naive estimates of $\mu_{\text{diff}}$ include (a) the complete-case (CC) estimate $\bar{y}_2 - \bar{y}_1$, where $\bar{y}_j = \sum_{i=1}^{N} Y_{ij}/r$ is the sample mean of $Y_j$ from the cases with both variables observed, and (b) the available-case (AC) estimate $\bar{y}_2 - \hat{\mu}_1$, where $\hat{\mu}_1 = \sum_{i=1}^{N} Y_{i1}/N$ is the sample mean of $Y_1$ from all the cases; the latter is obtained when the missing values of $Y_{i2}$ are imputed by the CC mean $\bar{y}_2$. We consider ML estimates for the following normal model: $$(Y_{i1}, Y_{i2}) \sim_{\text{ind}} N(\mu, \Sigma).$$

The MAR assumption (18.4) here implies that missingness of $Y_{i2}$ can depend on $Y_{i1}$, but conditional on $Y_{i1}$ it does not depend on $Y_{i2}$, since that variable is missing for $i = r + 1, \ldots, N$. The likelihood assuming ignorable non-response is

$$L_{\text{ign}}(\mu, \Sigma | \mathbf{y}) = \prod_{i=1}^{r} (\Sigma)^{-1/2} \exp \left\{ -0.5(\mathbf{y}_i - \mu)\Sigma^{-1}(\mathbf{y}_i - \mu) \right\}$$

$$\times \prod_{i=r+1}^{N} \sigma_{11}^{-1/2} \exp \left\{ -0.5(Y_{i1} - \mu_1)^2/\sigma_{11} \right\}. \quad (18.5)$$

Anderson (1957) showed that the ML estimates are easily derived by an elegant trick: instead of attempting to maximize (18.5) directly, the likelihood is factored according to the marginal distribution of $Y_{i1}$ and the conditional distribution of $Y_{i2}$ given $Y_{i1}$ (see also Little and Rubin, 2002, Chapter 7). The ML estimates of the marginal mean and variance
of $Y_1$ are sample means and variances from all $N$ cases:

$$\hat{\mu}_1 = \frac{1}{N} \sum_{i=1}^{N} y_{i1}, \quad \hat{\sigma}_{11} = \frac{1}{N} \sum_{i=1}^{N} (y_{i1} - \hat{\mu}_1)^2,$$

and the ML estimates of the slope, intercept, and residual variance of the regression of $Y_2$ on $Y_1$ are their least-squares estimates based on the $r$ complete cases. The corresponding ML estimates of $(\mu_2, \sigma_{12}, \sigma_{22})$ are:

$$\hat{\mu}_2 = \bar{y}_2 + \hat{\beta}_{21.1} (\hat{\mu}_1 - \bar{y}_1),$$

$$\hat{\sigma}_{12} = \hat{\beta}_{21.1} \hat{\sigma}_{11},$$

$$\hat{\sigma}_{22} = s_{22.1} + \hat{\beta}_{21.1} (\hat{\sigma}_{11} - s_{11}),$$

where $s_{jk} = \sum_{i=1}^{r} (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k)/r$ are sample variances ($j = k$) and covariances ($j \neq k$) from the $r$ complete cases, and $\hat{\beta}_{21.1} = s_{12}/s_{11}$ is the regression coefficient of $Y_1$ from the regression of $Y_{12}$ on $Y_{11}$, based on the complete cases. The ML estimate (18.6) of $\mu_2$ is called the regression estimate of the mean, and is also the average of observed and imputed values when the missing values of $y_2$ are imputed with predictions from the regression of $Y_{12}$ on $Y_{11}$ computed using the complete cases. The ML estimate of $\mu_{\text{diff}} = \mu_2 - \mu_1$ is

$$\hat{\mu}_{\text{diff}} = \hat{\mu}_2 - \hat{\mu}_1.$$

Large-sample inference requires estimates of standard errors of these parameters. Large-sample standard errors can be based on the observed information matrix, or can be computed by bootstrapping the observed data.

Another approach is to add a prior distribution and simulate draws from the posterior distribution of the parameters. With the non-informative prior

$$f(\mu_1, \sigma_{11}, \beta_{20.1}, \beta_{21.1}, \sigma_{22.1}) \propto \sigma_{11}^{-1} \sigma_{22.1}^{-1},$$

draws $(\mu_{1}^{(d)}, \sigma_{11}^{(d)}, \beta_{20.1}^{(d)}, \beta_{21.1}^{(d)}, \sigma_{22.1}^{(d)})$ from the posterior distribution of $(\mu_1, \sigma_{11}, \beta_{20.1}, \beta_{21.1}, \sigma_{22.1})$ are easily obtained as follows:

1. Draw independently $x_{1d}^2$ and $x_{2d}^2$ from chi-squared distributions with $N - 1$ and $r - 2$ degrees of freedom, respectively. Also, draw three standard normal deviates $z_{1d}, z_{2d},$ and $z_{3d}$.

2. Compute

$$\sigma_{11}^{(d)} = N \hat{\sigma}_{11} / x_{1d}^2,$$

$$\sigma_{22.1}^{(d)} = rs_{22.1} / x_{2d}^2,$$

$$\mu_{1}^{(d)} = \hat{\mu}_1 + z_{1d} \left( \sigma_{11}^{(d)} / N \right)^{1/2},$$

$$\beta_{21.1}^{(d)} = \hat{\beta}_{21.1} + z_{2d} \left( \sigma_{22.1}^{(d)} / rs_{11} \right)^{1/2},$$

$$\beta_{20.1}^{(d)} = \bar{y}_2 - \hat{\beta}_{21.1} \bar{y}_1 + z_{3d} \left( \sigma_{22.1}^{(d)} / r \right)^{1/2}.$$

Draws $(\mu_{2}^{(d)}, \sigma_{12}^{(d)}, \sigma_{22}^{(d)})$ from the posterior distribution of $(\mu_2, \sigma_{12}, \sigma_{22})$ are then obtained by replacing the ML estimates $(\hat{\mu}_1, \hat{\sigma}_{11}, \hat{\beta}_{20.1}, \hat{\beta}_{21.1}, \hat{\sigma}_{22.1})$ in (18.6) by the draws $(\mu_{1}^{(d)}, \sigma_{11}^{(d)}, \beta_{20.1}^{(d)}, \beta_{21.1}^{(d)}, \sigma_{22.1}^{(d)})$. Also $\mu_{\text{diff}}^{(d)} = \mu_{2}^{(d)} - \mu_{1}^{(d)}$ is a draw from the posterior distribution of $\mu_{\text{diff}}$.

For more details, see Little and Rubin (2002, Chapter 7).

This factored likelihood approach is readily extended to $n$ repeated measures $(Y_{11}, \ldots, Y_{in})$ with an $n$-variate $N(\mu, \Sigma)$ distribution, and a general monotone pattern with $Y_{ij}$ observed if $Y_{ik}$ is observed for all $j < k$. For an unrestricted mean and covariance matrix the factored likelihood idea leads to explicit expressions for ML estimates and draws from the posterior
distribution. Non-monotone patterns, or monotone patterns with more restricted parameterizations as in the next model, require iterative algorithms.

**Model 2. A normal mixed model for repeated measures with MAR missing data.** A model that is more tuned to repeated-measures data is the linear mixed model

\[
Y_i|b_i \sim N(X_i\beta + Z_i b_i, V_i)
\]

\[
b_i \sim N(0, G),
\]

where \(b_i\) are unobserved random effects for individual \(i\), and \(X_i\) and \(Z_i\) are known fixed matrices that characterize how the repeated measures depend on fixed and random factors in the model (Hartley and Rao, 1967; Laird and Ware, 1982; Jennrich and Schluchter, 1986; Schluchter, 1988). The matrices \(V_i\) and \(G\) characterize the covariance matrix of the repeated measures. Programs like `proc mixed` in SAS (Littell et al., 1996; SAS, 2004) and S-Plus (Pinheiro and Bates, 2000; Huet et al., 2004) include flexible choices of \(X_i, Z_i, V_i,\) and \(G,\) allowing a wide range of repeated-measures models to be fitted. Asymptotic inferences for this model are similar to the complete-data case.

Extensions to non-normal errors include models for multivariate \(t\) errors that downweight outliers (Lange, Little, and Taylor, 1989), and ML for generalized linear mixed models (McCulloch and Searle, 2001), which are implemented in `proc nlmixed` in SAS (2004). For non-linear mixed models, see Vonesh and Chinchilli (1997). The Bayesian approach is attractive in small samples where asymptotic assumptions are not appropriate. An early discussion of Bayesian methods for normal models is Gelfand et al. (1990). These methods can be implemented using the Bayesian modeling software in the BUGS project; see the BUGS Web site (BUGS, 2006) for details.

**Example 1. Longitudinal study of lung function.** Lavange and Helms (1983) analyzed data from a longitudinal study of lung function conducted on 72 children aged 3 to 12 years. A measure of maximum expiratory flow rate was measured annually, and differences in the resulting curve were related to between-subject covariates such as race and gender. The indexing variable of interest here is age rather than time. The number of actual measurements recorded on each child ranged from 1 to 8, with an average of 4.2 per child. A primary reason for the missing data was that children entered the study at different ages — some values were missing because the child was over age 3 at the start of the study, or less than age 12 at the end of the study. There was also some attrition from the study.

When, as here, the missing data are caused by features of the study design, rather than the behavior of the study subjects, the MAR assumption may be quite plausible. The missing-data mechanism depends on cohort, and if cohort is included as a covariate in the model, it is a special form of MAR, which we call covariate-dependent missingness. On the other hand, subjects who drop out prematurely may do so for reasons related to the outcome measures. For example, they may move out of the area because of respiratory problems. Such a mechanism is only MAR if the recorded values \(Y^o_i\) characterize the respiratory problems that prompted the move.

**Example 2. Dropouts in a hypertension trial.** Murray and Findlay (1988) described data from a large multicenter trial of metoprolol and ketanserin, two anti-hypertensive agents for patients with mild to moderate hypertension, with diastolic blood pressure an outcome measure of interest. The double-blind treatment phase lasted 12 weeks, with clinic visits scheduled for weeks 0, 2, 4, 8, and 12. The protocol stated that patients with diastolic blood pressure exceeding 110 mmHg at either the 4- or 8-week visit should “jump” to an open follow-up phase — a form of planned dropout. In total, 39 of the 218 metaprolol patients and 55 of the 211 ketanserin patients jumped to open follow-up. A further 17 metoprolol patients and 20 ketanserin patients had missing data for other reasons, of which
Analyses of the observed data indicate clearly that the dropouts differed systematically from cases remaining in the study, as can be predicted by the protocol for jumping to the open phase.

### 18.3 Non-ignorable models for the joint distribution of $Y$ and $R$

We now relax the assumption of ignorable non-response, and consider models for the joint distribution of $R_i$ and $Y_i$. These can be specified in a variety of ways, depending how the joint distribution is factorized. We first consider fixed-effect models that do not include random effects for subjects, and then consider mixed-effect models that use random-effect terms to model the longitudinal structure.

**Selection models** specify the joint distribution of $R_i$ and $Y_i$ through models for the marginal distribution of $Y_i$ and the conditional distribution of $R_i$ given $Y_i$:

$$f(R_i, Y_i | X_i, \gamma, \phi) = f_Y(Y_i | X_i, \gamma)f_{R|Y}(R_i | X_i, Y_i, \phi), \quad (18.8)$$

where $\theta = (\gamma, \phi)$. **Pattern-mixture models** (Glynn, Laird, and Rubin 1986; Little 1993) specify the marginal distribution of $R_i$ and the conditional distribution of $Y_i$ given $R_i$:

$$f(R_i, Y_i | X_i, \nu, \delta) = f_R(R_i | X_i, \delta)f_{Y|R}(Y_i | X_i, R_i, \nu), \quad (18.9)$$

where $\theta = (\nu, \delta)$. Applications of models of the form (18.9) for categorical outcomes include Ekholm and Skinner (1998) and Birmingham and Fitzmaurice (2002). Pattern-set mixture models (Little 1993), which are mixtures of these two types, can also be formulated.

For comparisons of selection and pattern-mixture models, see Glynn, Laird, and Rubin (1986), Little (1995), Kenward and Molenberghs (1999), and Michiels, Molenberghs, and Lipsitz (1999). Both of these modeling approaches have useful features. Attractive features of selection models include the following:

1. Selection models (18.8) are a natural way of factoring the model, with $f_Y$ the model for the data in the absence of missing values, and $f_{R|Y}$ the model for the missing-data mechanism that determines what parts of $Y$ are observed. Substantively it seems more natural to consider relationships between $Y$ and $X$ in the full target population of interest, rather than in subpopulations defined by missing-data pattern. In particular, the term $\gamma$ in the distribution $f_Y$ usually contains the parameters of substantive interest, and inferences for these parameters are available directly from the selection model analysis.

2. If the MAR assumption is plausible, the selection model formulation leads directly to the ignorable likelihood — the distribution $f_{R|Y}$ for the missing-data mechanism is not needed for likelihood inferences, which can be based solely on the model for $f_Y$. Thus, if MAR is viewed as reasonable, NMAR models are not contemplated, and inferences are required for the population aggregated over the missing-data patterns, then the selection modeling approach seems compelling, and I see little reason for considering a pattern-mixture formulation. For a discussion of MAR from a pattern-mixture model perspective, see Molenberghs et al. (1998).

Pattern-mixture models have some desirable features when NMAR situations are contemplated:

3. For situations where it is not substantively meaningful to consider non-response as missing data, it may make better sense to restrict the inference to the subpopulation of cases with values observed. For example, if $Y_{ij}$ is a measure of quality of life at age $j$, and $R_{ij} = 1$ for survivors at age $j$ and $R_{ij} = 0$ for individuals who die before age $j$, then it appears more meaningful to consider the distribution of $Y_{ij}$ given $R_{ij} = 1$ rather than the
marginal distribution of $Y_{ij}$, which effectively implies imputed quality-of-life measures for non-survivors. The pattern-mixture model formulation targets the distribution of substantive interest in this situation, and indeed a selection model that fails to condition on response is not sensible.

4. From an imputation perspective (see Section 17.4.1), missing values $Y_i^m$ should be imputed from their predictive distribution given the observed data including $R_i$, that is, $f(Y_i^m | Y_i^o, R_i, X_i)$. Under MAR this equals $f_Y(Y_i^m | Y_i^o, X_i)$, which is a conditional distribution derived from the selection model distribution of $Y$ given $X$. However, if data are not MAR, the predictive distribution of $Y_i^m$ given $Y_i^o$ and $R_i$ is modeled directly in the pattern-mixture formulation (18.9), but it is related to the components of the selection model by the complex expression

$$f(Y_i^m | Y_i^o, R_i, X_i) = \frac{f_Y(Y_i^m | Y_i^o, X_i)f_{R_i | Y}(R_i | X_i, Y_i)}{\int f_{R_i | Y}(R_i | X_i, Y_i)f_Y(Y_i^m | Y_i^o, X_i)dY_i^o}.$$ 

The more direct relationship between the pattern-mixture formulation and the predictive distribution for imputations yields gains in transparency and computational simplicity in some situations.

5. The selection model factorization does not require full specification of the model for the missing-data mechanism when the data are MAR, but it does if the data are NMAR. Some pattern-mixture models, such as Model 4 below, avoid specification of the model for the missing-data mechanism in NMAR situations, by using assumptions about the mechanism to yield restrictions on the model parameters.

18.4 Bivariate data with dropouts

We illustrate these general points with some examples of models, starting simple and then adding complexities.

**Model 3. A normal selection model for two repeated measures with non-MAR dropouts.** Suppose, as for model 1, there are just two repeated measures ($n_i = 2$ for all $i$), and $(Y_{i1}, Y_{i2})$ are observed and $R_i = 1$ for $i = 1, \ldots, r$, and $Y_{i2}$ is missing, and $R_i = 0$ for $i = r + 1, \ldots, N$. Let $\mu = (\mu_1, \mu_2)$, $\mu_j = E(Y_{ij})$ and $\Sigma = (\sigma_{jk})$ be the covariance matrix of $(Y_{i1}, Y_{i2})$, and suppose that interest lies in the difference in means $\mu_{dif} = \mu_2 - \mu_1$. We consider ML estimates for the following normal selection model:

$$(Y_{i1}, Y_{i2}) \sim N(\mu, \Sigma),$$  
$$(R_i | Y_{i1}, Y_{i2}) \sim \text{Ber}(P(\phi(Y_{i1}, Y_{i2}))),$$  
$$\logit\{P(\phi(Y_{i1}, Y_{i2}))\} = \phi_0 + \phi_1 Y_{i1} + \phi_2 Y_{i2},$$

where $\text{Ber}(\pi)$ represents the Bernoulli distribution with probability $\pi$. Replacing the logit specification for the response mechanism by a probit specification yields a simple case of the Heckman (1976) selection model. The likelihood for this model is

$$L(\mu, \Sigma, \phi | R, Y^o) = \prod_{i=1}^r [\Sigma]^{-1/2} \exp \left\{ -0.5(Y_i - \mu)' \Sigma^{-1} (Y_i - \mu) \right\} P(\phi(Y_{i1}, Y_{i2}))$$
$$\times \prod_{i=r+1}^N \sigma_{11}^{-1/2} \int \exp \left\{ -0.5(Y_i - \mu)' \Sigma^{-1} (Y_i - \mu) \right\} \{1 - P(\phi(Y_{i1}, Y_{i2}))\} dY_{i2}.$$ 

Joint ML estimation of $\gamma = (\mu, \Sigma)$ and $\phi = (\phi_0, \phi_1, \phi_2)$ requires an iterative method like the EM algorithm (Little and Rubin, 2002, Example 15.7). However, the model is weakly
identified, and identification is strongly dependent on the normality assumptions. Thus, preferred approaches are to make additional assumptions about the form of the mechanism, such as \( \phi_1 = 0 \) or \( \phi_2 = 0 \), or to do a sensitivity analysis by fitting the model for a variety of plausible choices of \( \phi \).

If \( \phi_2 = 0 \) then the missing data are MAR, since missingness of \( Y_{12} \) depends on \( Y_{11} \), which is observed for all \( i \). The full likelihood (18.11) then simplifies to

\[
L(\mu, \Sigma, \phi| R, Y^o) = L_{ign}(\mu, \Sigma| Y^o) \{ P[\phi(Y_{11}, Y_{12})] \}^T \{ 1 - P[\phi(Y_{11}, Y_{12})] \}^{N-r},
\]

where \( L_{ign}(\mu, \Sigma| Y^o) \) is given by (18.5), and ML inference for \( (\mu, \Sigma) \) can be based on the ignorable likelihood (18.5), as discussed in Model 1.

**Model 4. A normal pattern-mixture model for two repeated measures with non-MAR dropouts.** An unrestricted normal pattern-mixture model for the data just described is

\[
(Y_{11}, Y_{12}|R_i = k) \sim N(\mu^{(k)}, \Sigma^{(k)})
\]

\[
(R_i) \sim \text{Ber}(\delta).
\]

(18.12)

This model implies that the marginal mean of \( (Y_{11}, Y_{12}) \) averaged over patterns is \( \mu = (1-\delta)\mu^{(0)} + \delta \mu^{(1)} \), and the parameter of interest is

\[
\mu_{diff} = (1-\delta)(\mu^{(2)} - \mu^{(1)}) + \delta(\mu^{(2)} - \mu^{(1)}),
\]

(18.13)

the weighted average of the differences in means in the two patterns. Equation (18.13) is an example where the parameter of interest is not a parameter of the pattern-mixture model, but is easily expressed as a function of the model parameters; ML estimates or draws from the Bayesian posterior distribution are obtained by substituting ML estimates or Bayesian draws of the pattern-mixture model parameters in this expression.

The model (18.12) is clearly underidentified: there are 11 parameters, namely two means, two variances, and one covariance for the patterns of complete and incomplete cases, and the probability \( \delta \) that \( R_i = 1 \). On the other hand, only eight parameters \((\delta, \mu^{(1)}, \Sigma^{(1)}, \mu^{(0)}, \sigma^{(0)}_{11})\) can be estimated from the data. The ML estimates of these parameters are easily shown to be \( \tilde{\delta} = r/N, \tilde{\mu}^{(1)} = \bar{y}, \tilde{\Sigma}^{(1)} = S, \tilde{\mu}^{(0)} = \bar{y}^{(0)} \), and \( \tilde{\sigma}^{(0)}_{11} = s^{(0)}_{11} \), where \( \bar{y} = (\bar{y}_1, \bar{y}_2) \) and \( S = (s_{jk}) \) are the sample mean and covariance matrix of the complete cases, and \( \bar{y}^{(0)} \) and \( s^{(0)}_{11} \) are the sample mean and variance of \( Y_1 \) for the incomplete cases. The identification issue is more immediately evident with the model (18.12) than with the selection model (18.10), but it is a key issue regardless of how the joint distribution of \( Y \) and \( R \) is factorized.

Two possible resolutions of the identification issue are (a) to place restrictions on the model parameters, based on assumptions about the nature of the missing-data mechanism or the model for \( Y \), or (b) to relate the unidentified parameters to identified parameters via Bayesian prior distributions. A simple illustration of (a) is that if the missing-data mechanism is assumed missing completely at random, then \( \mu^{(1)} = \mu^{(0)} = \mu, \Sigma^{(1)} = \Sigma^{(0)} = \Sigma \), and model (18.12) is identical to the selection model (18.10) with \( \phi_1 = \phi_2 = 0 \). Likelihood inference for the parameters \((\mu, \Sigma)\) of interest is then the same as for the MAR model in Example 1 — it is not affected by the additional constraint that \( \phi_1 = 0 \).

Another restriction is to set \( \mu^{(0)} = \mu^{(1)} + \alpha, \Sigma^{(0)} = C\Sigma^{(1)}C' \) for pre-chosen values of \( \alpha \) and \( C \). That is, offsets are introduced to characterize differences in the mean and covariance matrix between complete and incomplete patterns (Daniels and Hogan, 2000). One possible approach is to assess sensitivity to pre-chosen values of \( \alpha \) and \( C \). A severe disadvantage of this strategy is that even for this simple missing-data pattern \( \alpha \) and \( C \) contain five distinct quantities, and prespecifying values in a five-dimensional space seems impractical, and unnecessary since only three parameter restrictions are needed to identify
the model. Daniels and Hogan note that some of these parameters can be estimated from the
data, providing assumptions are made about the missing-data mechanism. They describe
sensitivity analyses based on this approach on longitudinal clinical trial data involving
growth hormone treatments.

Little (1994) analyzes model (18.12) under the assumption that
\[ \Pr(R_i = 0 | Y_{i1}, Y_{i2}) = g(Y_{i1}^*), \quad Y_{i1}^* = Y_{i1} + \lambda Y_{i2}, \]
(18.14)
where \( \lambda \) is unspecified and \( g \) is an arbitrary function. Under that assumption, the conditional
distribution of \( Y_{i2} \) given \( Y_{i1}^* \) is independent of \( R_i \), and is normal with, say, mean \( \beta_{20} + \beta_{21} Y_{i1}^* \) and variance \( \sigma_{22}^2 \). The fact that the intercept, slope, and residual variance of this
distribution are the same in the two patterns yields three constraints on \( (\mu^{(0)}, \Sigma^{(0)}) \) that are just sufficient to identify the model. Little (1994) shows that the resulting ML estimates
of \( (\mu_1, \mu_2) \) are \( \hat{\mu}_1 = \sum_{i=1}^N y_{i1}/N \), and
\[ \hat{\mu}_2^{(\lambda)} = \bar{y}_2 + \hat{\beta}_{21}^{(\lambda)} (\hat{\mu}_1 - \bar{y}_1), \quad \hat{\beta}_{21}^{(\lambda)} = \frac{s_{12} + \lambda s_{22}}{s_{11} + \lambda s_{12}}. \]
The corresponding estimate of the difference in means is
\[ \hat{\mu}_{\text{diff}}^{(\lambda)} = \hat{\mu}_2^{(\lambda)} - \hat{\mu}_1 \]
(18.15)
\[ = \bar{y}_2 - ((\hat{\beta}_{21}^{(\lambda)} + (1 - \hat{\beta}_{21}^{(\lambda)})(r/N))\bar{y}_1 + (1 - r/N)(1 - \hat{\beta}_{21}^{(\lambda)}))\bar{y}_1, \]
where \( \bar{y}_1^{(0)} \) is the mean of \( Y_1 \) for the cases missing \( Y_2 \). Various estimates \( \hat{\mu}_{\text{diff}}^{(\lambda)} \) are obtained for
different choices of \( \lambda \), including the CC and AC estimates previously mentioned in Model 1.
The following comments assume \( Y_1 \) and \( Y_2 \) are positively correlated for the complete cases:

(a) When \( \lambda = 0 \) the data are MAR, and \( \hat{\mu}_{\text{diff}}^{(\lambda=0)} = \hat{\mu}_{\text{diff}} \), the ML estimate for the ignorable
selection model.

(b) As \( \lambda \) increases from 0 to \( (s_{11} - s_{12})/(s_{22} - s_{12}) \), the ML estimate puts less and less
weight on \( \bar{y}_1^{(0)} \), the mean of \( Y_1 \) for the cases that drop out. When \( \lambda = (s_{11} - s_{12})/(s_{22} - s_{12}), \hat{\beta}_{21}^{(\lambda)} = 1 \) and \( \hat{\mu}_{\text{diff}}^{(\lambda)} = \bar{y}_2 - \bar{y}_1 \), the CC estimate. This value of \( \lambda \) reduces to \( \lambda = 1 \)
when \( s_{11} = s_{22} \). The implication is that when missingness depends on the average value
of the \( Y \) at the two time points, the CC estimate of the change is optimal.

(c) As \( \lambda \) increases from \( (s_{11} - s_{12})/(s_{22} - s_{12}) \) toward \( \infty \), \( \hat{\beta}_{21}^{(\lambda)} \to s_{22}/s_{12}, \) and the slope
of the regression predictions of missing values of \( Y_2 \) is the inverse of the slope of the
regression of \( Y_1 \) on \( Y_2 \). This calibration-like estimate reverses the regressions, and is
increasingly unstable as the correlation between \( Y_1 \) and \( Y_2 \) tends to zero.

(d) As \( \lambda \) decreases from 0 to \(-s_{12}/s_{22}, \) the ML estimate puts increasing weight on \( \bar{y}_1^{(0)} \).
When \( \lambda = -s_{12}/s_{22}, \hat{\mu}_{\text{diff}}^{(\lambda)} = \bar{y}_2 - \bar{y}_1 \), the AC estimate.

(e) When \( \lambda = -s_{11}/s_{12}, \) the estimate \( \hat{\mu}_{\text{diff}}^{(\lambda)} \) is indeterminate since \( \hat{\beta}_{21}^{(\lambda)} \) has a denominator
equal to zero.

(f) Suppose \( s_{11} = s_{22} \) and the sample correlation for the complete cases is \( r \). Then \( \lambda = -r \)
leads to the AC estimate \( \bar{y}_2 - \bar{y}_1 \), and \( \lambda = -1/r \) leads to an indeterminate estimate.
These two possibilities bracket \( \lambda = -1 \), where missingness depends on the change
\( Y_2 - Y_1 \).

This relatively simple example illustrates the role of the missing-data mechanism in the
properties of estimates. It is tempting to attempt to estimate \( \lambda \) from the data, but unfortunately
the various choices of \( \lambda \) noted above all give the same fits to the observed data.
Hence, we are reduced to a sensitivity analysis where \( \lambda \) is varied over a plausible range. This
approach is illustrated for a more complex model in Example 3 to follow.
Model 5. Selection model for \( n \) repeated measures with non-MAR dropouts and covariates. The previous two models are easily extended to data on \( n \) repeated measures \( Y_i = (Y_{i1}, \ldots, Y_{in}) \) with the last measure \( Y_{in} \) subject to dropout, and \( R_i = 1 \) if \( Y_{in} \) is observed, and \( R_i = 0 \) if \( Y_{in} \) is missing. We also include a set of fully observed covariates \( X_i \), and for simplicity assume linear relationships between the repeated measures and these covariates. An NMAR selection model that extends model 3 is

\[
(Y_i \mid X_i) \sim N(X_i \beta, \Sigma),
\]

\[
(R_i \mid Y_i, X_i) \sim \text{Ber}(P(\phi(Y_i, X_i))),
\]

\[
\text{logit}\{P(\phi(Y_i, X_i))\} = \phi_0 + \phi_1 Y_{i1} + \ldots + \phi_{n-1} Y_{in-1} + \phi_n Y_{in} + \phi_{n+1} X_i.
\]

This model allows the response mechanism to depend on the values of \( Y_{i1}, \ldots, Y_{in} \) as well as the covariates. The MAR model corresponds to \( \phi_n = 0 \). The response propensity model could also be extended to allow interactions between the covariates and observed components of \( Y_i \).

Model 6. Pattern-mixture model for \( n \) repeated measures with non-MAR dropouts and covariates. A pattern-mixture analog of Model 5 is

\[
(Y_i \mid X_i, R_i = k) \sim N(X_i \beta^{(k)}, \Sigma^{(k)}),
\]

\[
(R_i \mid X_i) \sim \text{Ber}(\delta(X_i))\] where the parameters \((\beta^{(k)}, \Sigma^{(k)})\) of the multivariate normal distribution are different for each pattern. The following example applies this model.

Example 3. A dose-comparison study for schizophrenia treatments. Little and Wang (1996) used a model of the form (18.16) to analyze data from a clinical trial to compare three alternative dose regimens of haloperidol for schizophrenia patients. Sixty-five patients with DSM-III diagnosis of schizophrenia were assigned to receive 5, 10, or 20 mg/day of haloperidol for 4 weeks. The outcome variable \( Y \) was the Brief Psychiatric Rating Scale Schizophrenia (BPRSS) factor, measured at \( j = 1 \) (baseline), \( j = 2 \) (week 1), and \( j = 3 \) (week 4). The main parameters of interest were the average change in BPRSS between baseline and week 4 for each dose group. Twenty-nine patients dropped out of the study at \( j = 3 \), with dropout rates varying across dose groups. Accordingly, \( R_i = 1 \) if \( Y_{i3} \) is observed and \( R_i = 0 \) if \( Y_{i3} \) is missing. A poor BPRSS outcome may cause patients to leave the study, particularly if combined with unpleasant side effects associated with the drug, particularly at high doses. Thus, models were fit where missingness of BPRSS at week 4 depended not only on the dosage, but also on the BPRSS values at week 4 and at previous times. Little and Wang fitted the following pattern-mixture model:

\[
(Y_i \mid X_i, R_i = k) \sim N(X_i \beta^{(k)}, \Sigma^{(k)}),
\]

\[
(R_i \mid X_i) \sim \text{MNOM}(\pi(X_i, \delta)),
\]

\[
\text{logit}(\pi(X_i, \delta)) = \delta' X_i,
\]

where \( X_i \) represents three treatment dummies, and MNOM denotes the multinomial distribution. Thus, for pattern \( k \), \( Y_i \) has a trivariate normal linear regression on \( X_i \) with \((3 \times 3)\) coefficient matrix \( B^{(k)} \) and covariance matrix \( \Sigma^{(k)} \). The parameters of (18.17) are the \((3 \times 1)\) vector \( \delta \), estimated by the vector of observed non-response rates at week 4 for each dose group, and \((\beta^{(k)}, \Sigma^{(k)})\) for \( k = 0, 1 \).

This model is underidentified, in that there are no data to estimate directly the six parameters of the distribution of \( Y_{i3} \) given \( Y_{i1}, Y_{i2}, \) and \( X_i \) for the dropout pattern \( R_i = 0 \). These parameters are identified by assumptions about the missing-data mechanism. Specifically,
suppose it is assumed that
\[ \text{Pr}(R_i = 1) = g(c_{i1}Y_{i1} + c_{i2}Y_{i2} + c_{i3}Y_{i3}, X_i), \]  
(18.18)
where \( g \) is an arbitrary unspecified function, and \( c_i = (c_{i1}, c_{i2}, c_{i3}) \) are prespecified coefficients. When \( c_{i3} = 0 \) in (18.18), the conditional distribution of \( Y_{i3} \) given \((Y_{i1}, Y_{i2}, X_i)\) is the same for the complete and incomplete cases, and the data are MAR. The effect of non-ignorable non-response was assessed by computing ML and Bayes estimates for various other choices of \( c_i \).

Specifically, Table 18.1 shows estimates of the difference in mean BPRSS between baseline and week 4 for the three treatment groups, for the following methods:

1. CC analysis, where incomplete cases are dropped from the analysis.
2. Ignorable ML, where missingness is assumed to depend on the BPRSS scores at baseline and week 1. These results are ML under the ignorable pattern-mixture model or the ignorable selection model. Standard errors are the standard deviation of estimates from 1000 bootstrap samples.
3. ML under the pattern-mixture model (18.17) and (18.18), with the following alternative choices of \( c_i \): A. \( c_i = (0.4, 0.4, 0.2) \); B. \( c_i = (0.3, 0.4, 0.4) \); C. \( c_i = (0.1, 0.1, 0.8) \); and D. \( c_i = (0, 0, 1) \). These represent progressively more extreme departures from MAR, with A being closest to the ignorable assumption \( c_{i3} = 0 \) corresponding to method (2). ML estimates were computed using an EM algorithm, and asymptotic standard errors were computed using the SEM algorithm (Meng and Rubin, 1991), which provides a numerical approximation to the inverse of the observed covariance matrix.
4. For each of the models in (3), the mean and variance of the posterior distribution of the parameters, based on a non-informative prior. The posterior distributions were simulated by Gibbs sampling.

It can be seen from Table 18.1 that (a) the CC estimates deviate noticeably from estimates from the other methods, a common finding when the amount of missing data is substantial; (b) the ML/SEM and Bayes estimates for the pattern-mixture models are broadly similar, and the asymptotic standard errors are somewhat smaller than the posterior standard errors, particularly for pattern D; the posterior standard errors are preferred because they do not assume large samples; (c) the size of treatment effects is only moderately sensitive to the choice of model (ignorable, A–D); the effect of choice of model is more pronounced in the high-dose group than in the other groups, reflecting the higher dropout

| Table 18.1 Example 3: Estimates (Standard Errors) of Differences of Means of BPRSS between Baseline and Week 4, under Various Models |
|---------------------------------|-----------------|-----------------|-----------------|
| Methods                         | Dose 5          | Dose 10         | Dose 20         |
| (1) Complete Cases              | 3.70 (1.03)     | 4.35 (0.73)     | 5.67 (1.33)     |
| (2) Ignorable ML                | 3.29 (0.90)     | 4.09 (0.62)     | 6.46 (1.04)     |
| (3) Pattern-Mixture Models: ML  |                 |                 |                 |
| Mechanism A                     | 3.28 (0.90)     | 4.14 (0.62)     | 6.53 (1.05)     |
| Mechanism B                     | 3.25 (0.91)     | 4.18 (0.63)     | 6.61 (1.07)     |
| Mechanism C                     | 3.18 (0.95)     | 4.25 (0.66)     | 6.81 (1.16)     |
| Mechanism D                     | 3.14 (0.97)     | 4.27 (0.68)     | 6.91 (1.21)     |
| (4) Pattern-Mixture Models: Bayes|                |                 |                 |
| Mechanism A                     | 3.23 (0.99)     | 4.07 (0.71)     | 6.46 (1.19)     |
| Mechanism B                     | 3.21 (1.02)     | 4.13 (0.72)     | 6.56 (1.22)     |
| Mechanism C                     | 3.13 (1.12)     | 4.23 (0.77)     | 6.81 (1.39)     |
| Mechanism D                     | 3.08 (1.19)     | 4.26 (0.82)     | 6.96 (1.53)     |
rate for that group; (d) as missingness becomes increasingly dependent on the missing week 4 BPRSS value, the mean treatment effects decrease slightly for the low- and moderate-dose groups, and increase somewhat more for the high-dose group. The net effect of this change in assumed mechanism is to slightly increase the differentials in treatment effects by size of dose; (e) the standard errors of the pattern-mixture model estimates increase from models A through D, reflecting a loss of information with increasing degree of non-ignorable non-response.

18.5 Mixed models with dropouts

The scope of models for MAR repeated-measures data was expanded in Model 2 by including unobserved within-subject random effects \( b_i \) in the model. With NMAR data, the selection and pattern-mixture formulations can be expanded to allow the possibility that the missing-data mechanism depends on these random effects (Little, 1995). This leads to a rich class of models based on various factorizations of the joint distribution of \( R_i, Y_i, \) and \( b_i \). There are six ways this joint distribution can be factored, and three of them condition the distribution of \( b_i \) on \( Y_i \), which is not sensible here. The remaining three factorizations yield mixed-effect selection models of the form

\[
f(R_i, Y_i, b_i | X_i, \gamma) = f_B(b_i | X_i, \gamma_1)f_{Y_i|B}(Y_i | X_i, b_i, \gamma_2)f_{R_i|Y,B}(R_i | X_i, Y_i, b_i, \gamma); \quad (18.19)
\]

mixed-effect pattern-mixture models of the form

\[
f(R_i, Y_i, b_i | X_i, \delta, \nu) = f_R(R_i | X_i, \delta)f_{B|X,Y}(b_i | X_i, \nu_1)f_{Y_i|B}(Y_i | X_i, b_i, \nu_2); \quad (18.20)
\]

and mixed-effect hybrid models of the form

\[
f(R_i, Y_i, b_i | X_i, \delta, \nu) = f_R(R_i | X_i, \delta)f_{B|X,Y}(b_i | X_i, \nu_1)f_{Y_i|B,R}(Y_i | X_i, b_i, R_i, \nu_2). \quad (18.21)
\]

Models based on (18.21) have not, to my knowledge, been considered in the literature. I consider some examples of (18.19) and (18.20), focusing on various assumptions about the missing-data mechanism.

18.5.1 Covariate-dependent dropout

The two factorizations (18.19) and (18.20) become equivalent under the strong assumption that the dropout mechanism does not depend on outcome values \( Y_i \) or the random effects \( b_i \), but depends only on the values of fixed covariates \( X_i \), that is, for (18.19),

\[
f_{R_i|Y,B}(R_i | X_i, Y_i, b_i, \phi) = f_R(R_i | X_i, \phi) \quad (18.22)
\]

This model is a strong form of MAR that allows dependence of dropout on both between-subject and within-subject covariates that can be treated as fixed in the model. In particular, dropout can depend on treatment-group indicators or other baseline covariates that are included in the model. Diggle and Kenward (1994) called assumption (18.22) “completely random dropout,” and viewed it as a special case of Rubin’s (1976) missing completely at random assumption. Little (1995) reserves the term “missing completely at random” for the case when missingness does not depend on \( X_i \) as well as \( Y_i \) and \( b_i \). Assumption (18.22) is capable of some empirical verification, by comparing empirical distributions of observed outcomes \( Y_i^o \) across patterns after adjusting for the covariates. For example, if there are two outcomes, and \( Y_{1i} \) is fully observed and \( Y_{12} \) has missing values, then one can compare the adjusted mean of \( Y_{1i} \) given \( X_i \) for the complete and incomplete cases. One way of implementing this is to regress \( Y_{1i} \) on \( X_i \) and the indicator \( R_{i2} \) for whether \( Y_{12} \) is missing, and test whether the coefficient of \( R_{i2} \) is different from zero.

Under covariate-dependent missingness, analysis of the complete cases is not biased, although it is subject to a loss of efficiency — indeed in Example 1 above it is not feasible,
since the design of the study results in no complete cases. Methods that use all the data, such as ML or GEE, are generally more efficient and yield consistent estimators under the usual assumptions of these methods.

18.5.2 MAR dropout

The MAR assumption in the context of (18.19) assumes that dropout depends on \( Y_i \) and \( b_i \) only through the observed data \( Y_i^o \), that is,

\[
f_{R|Y,B}(R_i|X_i, Y_i, b_i, \phi) = f_{R|Y}(R_i|X_i, Y_i^o, \phi).
\]

The clinical trial of two anti-hypertensive agents in Example 2 provides an illustration of a case where MAR is plausible, since dropout, namely moving to the open phase of the protocol, depends on a blood pressure value that is recorded. Although MAR, the ML analysis is vulnerable to misspecification of the relationship between outcomes and blood pressure because this can only be estimated for the recorded blood pressures, which are lower than the blood pressures for cases after dropout. Dropping-out because of side effects would also be MAR if the side effects were recorded and included in the analysis via likelihood methods. As noted in comment (ii) in the Introduction, multiple imputation based on a model that includes the side-effect data is perhaps the most convenient approach for achieving this. This was not done in Murray and Findlay (1988), but the number of such cases is small. The nature of the mechanisms for the 37 patients who dropped out for “other reasons” is not discussed, but at least the predominant dropout mechanism here is plausibly MAR.

Under MAR and the distinctness condition noted above, the missing-data mechanism is ignorable, and ML or Bayes inference can be based on models such as Model 2. In contrast, other methods such as CC analysis or GEE generally require the stronger assumption (18.22) to yield consistent estimators (Fitzmaurice, Laird, and Rotnitzky 1993; Kenward, Lesaffre, and Molenberghs 1994). This is an important advantage of likelihood-based inference, although the methods require adequate specification of the model. The GEE approach can be modified to yield consistent estimators under the MAR assumption, by multiplying GEE weights by the inverse of estimated selection probabilities. (Robins, Rotnitsky, and Zhao, 1995). For a related use of estimated selection probabilities to adjust for random dropout in a simpler situation, see Heyting, Tolboom, and Essers (1992).

18.5.3 Non-ignorable outcome-dependent dropout

In other settings, dropout may depend on missing components \( Y_i^m \) of \( Y_i \), such as the (unrecorded) value of the outcome at the time when the subject drops out. Little (1995) calls the resulting assumption “outcome-dependent dropout”:

\[
f_{R|Y,B}(R_i|X_i, Y_i, b_i, \phi) = f_{R|Y}(R_i|X_i, Y_i^o, Y_i^m, \phi).
\]  

(18.23)

Diggle and Kenward (1994) used a model of the form (18.23) to analyze data from a longitudinal milk protein trial. Cows were randomly allocated to one of three diets (barley, mixed barley and lupins, and lupins) and assayed the protein content of milk samples taken weekly for a period of 20 weeks. Dropout corresponded to cows that stopped producing milk before the end of the experiment. The complete-data model \( f_{Y|B}f_B \) specified a quadratic model for the mean protein content over time, with an intercept that depended on diet (thus modeling an additive effect of treatment). The covariance structure was assumed to be a combination of an autoregressive structure with an added independent measurement error. The dropout distribution \( f_{R|Y} \) process was modeled as depending on the current and previous value of protein content, specifically:

\[
\logit\{\Pr(R_{it} = 1|R_{it-1} = 1, Y_i, X_i, \phi)\} = \phi_0 + \phi_1 Y_{it-1} + \phi_2 Y_{it}.
\]  

(18.24)
The resulting ML estimates $\hat{\phi}_1 = 12.0, \hat{\phi}_2 = -20.4$ of the coefficients $\phi_1$ and $\phi_2$ suggested that the probability of dropout increases when the prevailing level of protein is low or the increment between the last and current protein content is high. As noted in the discussion of the article, underidentifiability is a serious problem with this model. A controversial issue concerns whether the parameters in (18.24) can be simultaneously estimated with the parameters of the distributions of $f_B$ and $f_{Y|B}$. A sensitivity analysis might consider ML estimates of the parameters for a variety of plausible alternative choices of $\phi_1$ and $\phi_2$.

Models of the form (18.23) have been considered for non-normal data. In particular for repeated-measures ordinal data, Molenberghs, Kenward, and Lesaffre (1997) combined a multivariate Dale model for the outcomes with a model for the dropout mechanism analogous to that of Diggle and Kenward (1994). Problems of identification arise for these models too.

### 18.5.4 Non-ignorable random-coefficient dependent dropout

Another form of non-ignorable dropout model assumes dropout at time $t$ depends on the value of $b_i$, that is,

$$f_{R|Y,B}(R_i|X_i, Y_i, b_i, \phi) = f_{R|B}(R_i|X_i, b_i, \phi). (18.25)$$

Examples of models of dropout of the form (18.25) include Wu and Carroll (1988), Shih, Quan, and Chang (1994), Mori, Woolson, and Woodworth (1994), Schluchter (1992), and the following:

**Example 4. Longitudinal AIDS data.** DeGruttola and Tu (1994) modeled the relationship between the progression of CD4 lymphocyte count and survival for patients enrolled in a clinical trial of two alternative doses for zidovudine. Here, a vector of log CD4 counts for subject $i$ is modeled via a mixed model of the form (18.7). The main cause of dropout is death, which is measured as survival time, and modeled as a continuous, normally distributed random variable with a mean that is a linear function of covariates. ML estimation is accomplished using an EM algorithm (Dempster, Laird, and Rubin 1977), with standard errors computed using the method of Louis (1982). A drawback with this approach is that the selection-model factorization effectively treats the CD4 counts after death as missing values, which is not in accord with the definition of missing data provided above. In my view, a better analysis would condition the analysis of CD4 counts at any time on individuals who have survived up to that time.

### 18.5.5 Shared-parameter models

A number of models have been formulated that assume both the outcome process and the dropout process depend on shared latent variables. These are called *shared-parameter models*, and examples include Ten Have et al. (1998, 2002), Albert et al. (2002), and Roy (2003). They are special cases of (18.19) and (18.21) where $Y_i$ and $R_i$ are assumed independent given $b_i$:

$$f(R_i, Y_i, b_i|X_i, \gamma, \phi) = f_B(b_i|X_i, \gamma_1)f_{Y|B}(Y_i|X_i, b_i, \gamma_2)f_{R|B}(R_i|X_i, b_i, \phi).$$

**Example 5. A shared-parameter model for heroin addiction treatment data with missing data.** Albert et al. (2002) analyzed data from a clinical trial of treatments of heroin addiction that randomized patients into one of two treatment groups: buprenorphine ($n = 53$) and methadone ($n = 55$). Patients were scheduled for urine tests three times a week on Monday, Wednesday, and Friday for 17 weeks post-randomization (51 scheduled
MIXED MODELS WITH DROPOUTS

responses). Urine tests were scored as positive or negative for the presence of opiates at each follow-up visit. A primary scientific objective of the study was to compare the marginal proportion of positive urine tests over follow-up between the two treatment arms. Plots suggested that the frequency of positive urine tests was relatively constant over follow-up, with the buprenorphine arm having a lower proportion of positive urine tests than the methadone arm. Thus, the analyses focused on comparing the marginal proportions, assumed constant over time, across the two treatment arms.

The analysis was complicated by the unequally spaced observations and the large amount of missing data, which took the form of dropouts and intermittent missing data. A number of subjects withdrew from the study due to poor compliance or because they were offered places in treatment programs that gave unmasked treatment and long-term care. Intermittent missingness may be more closely associated with the response process, as patients may not show up when they are taking opiates. The missing-data mechanism appeared different in the two treatment arms. The proportion of patients dropping out by the end of the 17-week period was 80% in the methadone group and 59% in the buprenorphine group. In addition, patients had a sizable amount of intermittent missing data, with the proportion of patients dropping out by the end of the 17-week period being higher in the buprenorphine arm than the methadone arm. The Spearman rank correlation between the proportion of positive tests and the time to dropout was −0.44 in the buprenorphine arm and −0.10 in the methadone arm. The correlations between the proportion of positive tests and the proportion of intermittent missing visits before dropout in the buprenorphine and methadone arms were 0.40 and 0.29, respectively.

These calculations suggest that addicts who are more likely to use drugs are both more likely to dropout and to have a higher frequency of intermittent missing data before dropout than addicts who use opiates less frequently. These associations are consistent with an NMAR missing-data mechanism. The differences in the magnitude of these correlations between treatment arms suggest that the informative missing data may be greater in the buprenorphine arm than in the methadone arm.

For the ith patient, let \( y_{it1}, y_{it2}, \ldots, y_{itn} \) be the sequence of \( n \) intended binary measurements, where \( t_j \) is the time of the \( j \)th follow-up, and let \( Y_{it}^\circ \) be the vector of observed binary responses. Denote \( R_i = (R_{it1}, R_{it2}, \ldots, R_{itn}) \) as indicators of intermittent missingness or dropout at each follow-up time, where \( R_{it} = 0 \) if \( y_{it} \) is observed, \( R_{it} = 1 \) if \( y_{it} \) is intermittently missing, and \( R_{it} = 2 \) if \( y_{it} \) is a value after dropout. Both \( Y_i \) and \( R_i \) are modeled conditional on a latent process, \( \{b_{it}\} \), and a time-dependent covariate vector, \( \{X_{it}\} \). Specifically, a shared-parameter model of the following form is assumed (notation differs slightly from previous examples):

\[
\text{logit}\{\Pr(y_{itj} = 1|b_{itj})\} = X_{itj}'\beta + b_{itj},
\]

\[
\Pr(R_{itj} = \ell|b_{itj}, R_{itj-1} \neq \ell) = \begin{cases} 
\frac{1}{1 + \sum_{\ell=1}^{2} \exp(\nu_{itj}^\ell_1 \eta_1 + \psi b_{itj})}, & \ell = 0, \\
\frac{\exp(\nu_{itj}^\ell_1 \eta_1 + \psi b_{itj})}{1 + \sum_{\ell=1}^{2} \exp(\nu_{itj}^\ell_1 \eta_1 + \psi b_{itj})}, & \ell = 1, 2,
\end{cases}
\]

where \( \nu_{itj}^\ell_1 \) are vectors of covariates and \( \eta_1 \) their corresponding regression coefficients, and \( \psi \) are parameters that relate the missingsness (intermittent missing and dropout) with the outcome data. Since \( R_{it} = 2 \) denotes dropout, which is an absorbing state, \( \Pr(R_{itj+1} = 2|R_{itj} = 2) = 1 \). The shared random parameters \( \{b_{itj}\} \) are modeled as a Gaussian Ornstein–Uhlenbeck process (Feller, 1971), with mean zero and covariance

\[
\text{Cov}(b_{itj}, b_{it'}) = \sigma^2 \exp(-\theta|t - t'|), \text{ where } \theta > 0.
\]
In the opiate clinical trial application, a version of (18.26) with no time effects was fitted separately in the two treatment groups, namely:
\[
\logit\{\Pr(Y_{itj} = 1 | b_{itj})\} = \beta_G + b_{itj},
\]
\[
\Pr(R_{itj} = \ell | b_{itj}, R_{itj-1} \neq 2) = \begin{cases} 
1 + \sum_{\ell'=1}^{\ell} \exp(\theta_G + \gamma_G b_{itj}), & \ell = 0, \\
\exp(\eta_G + \gamma_G b_{itj}), & \ell = 1, 2,
\end{cases}
\]
\[
\Cov(b_{itj}, b_{it'}') = \sigma_G^2 \exp\left(-\theta_G |t - t'|\right), \quad \text{where } \theta > 0,
\]
where \(G = 0\) and \(G = 1\) index parameters in the methadone and buprenorphine groups, respectively. ML estimation was accomplished using a Monte Carlo EM algorithm, and standard errors estimated by the bootstrap with 250 bootstrap samples. Table 18.2 shows the resulting parameter estimates and standard errors. The parameter estimates for the buprenorphine group show a significant positive relationship between response, dropout, and intermittent missingness (estimates of \(\psi_{11}\) and \(\psi_{21}\) were highly significant), suggesting that the missing-data mechanism is non-ignorable for this group. The corresponding estimates for the methadone group were smaller in magnitude and not statistically significant at the 0.05 level, although they were positive.

### 18.6 Mixed-effect pattern-mixture models

Mixed-effect pattern-mixture models are based on the factorization (18.20). This approach stratifies the sample by pattern of missing data (e.g., by the time of dropout) and then models differences in the distribution of \(Y_i\) over these patterns. As with the selection models, these models can be formulated for the case where missingness depends on \(Y_i\),
\[
f(R_i, Y_i, b_i | X_i, \delta, \nu) = f_R(R_i | X_i, \delta) f_B(b_i | X_i, \nu_1) f_{Y|R}(Y_i | X_i, R_i, b_i, Y_i),
\]
and for the case where missingness depends on \(b_i\),
\[
f(R_i, Y_i, b_i | X_i, \delta, \nu) = f_R(R_i | X_i, \delta) f_B(b_i | X_i, R_i, \nu_1) f_{Y|R}(Y_i | X_i, b_i, \nu_2).
\]

The latter models have the computational advantage that parameters of the distribution \(f_{Y|R}(Y_i | X_i, R_i, b_i, \nu_2)\) can be estimated using existing mixed-model software, such as SAS proc mixed, by including the dropout indicators \(R_i\) as covariates in the model.

### Table 18.2 Parameter Estimates and Standard Errors (SE) for Shared-Parameter Model Fitted to Heroin Addiction Treatment Data of Example 5

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Methadone</th>
<th>SE</th>
<th>Buprenorphine</th>
<th>Estimate</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\beta_G)</td>
<td>1.44</td>
<td>0.43</td>
<td>-0.10</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>(\nu_{1G})</td>
<td>-1.89</td>
<td>0.22</td>
<td>-1.71</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>(\nu_{2G})</td>
<td>-3.42</td>
<td>0.21</td>
<td>-4.11</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>(\psi_{1G})</td>
<td>0.29</td>
<td>0.13</td>
<td>0.43</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>(\psi_{2G})</td>
<td>0.22</td>
<td>0.15</td>
<td>0.48</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>(\sigma_G)</td>
<td>2.84</td>
<td>0.55</td>
<td>2.77</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>(\theta_G)</td>
<td>0.014</td>
<td>0.008</td>
<td>0.012</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td>(P_G(R_{it} = 1))</td>
<td>0.67</td>
<td>0.040</td>
<td>0.49</td>
<td>0.039</td>
<td></td>
</tr>
<tr>
<td>(P_G(R_{it} = 1))</td>
<td>0.13</td>
<td>0.023</td>
<td>0.15</td>
<td>0.023</td>
<td></td>
</tr>
<tr>
<td>(P_G(R_{it} = 2))</td>
<td>0.027</td>
<td>0.005</td>
<td>0.014</td>
<td>0.004</td>
<td></td>
</tr>
</tbody>
</table>

Source: Albert et al. (2002).
MIXED-EFFECT PATTERN-MIXTURE MODELS

Fitzmaurice, Laird, and Schneyer (2001) discussed forms of (18.27) that are parameterized to enhance interpretability of the parameters, and apply it to data from an asthma trial with normal repeated measures. Hogan, Lin, and Herman (2004) apply a model of this form to AIDS clinical trial data, where the fixed-effects parameters in a mixed model for $Y_i$ given $X_i$ and $b_i$ are allowed to depend non-parametrically on the dropout time, which may be categorical or continuous. Two examples of models of this type are now presented:

**Model 7. A straight-line pattern-mixture model.** Suppose that $X_i$ is scalar (e.g., time or age), and

\[
\begin{align*}
(Y_{ij}|X_i, b_i, R_i = k, \nu_2) &\sim N(b_{0i} + b_{1i}X_i, \sigma^2), \\
(b_{0i}, b_{1i}|\nu_1) &\sim N((b_{0i}^{(k)}, b_{1i}^{(k)})^T, \Gamma), \\
Pr(R_i = k) &\pi_k,
\end{align*}
\]

which models $\{Y_{ij}\}$ with a linear regression on $X_i$ with random slope $b_{1i}$ and intercept $b_{0i}$, which are in turn distributed about a line with the same intercept $b_{0i}^{(k)}$ and slope $b_{1i}^{(k)}$ for each pattern $k$. This can be modeled via a standard mixed model by including as covariates dummy variables for each pattern. At least two repeated measures are needed to allow estimation of the slope and intercept for each pattern. If the quantities of interest are the expected intercept and slope, averaged over missing-data pattern, that is, $b_0 = \sum_{k=1}^K \pi_k b_{0i}^{(k)}$ and $b_1 = \sum_{k=1}^K \pi_k b_{1i}^{(k)}$, ML estimates of these parameters are obtained as a weighted sum of the ML estimates of the expected intercept and slope for pattern $k$, with weights given as the proportion of cases with pattern $k$. This contrasts with an MAR model, where estimates for each pattern are effectively weighted by their precision. This model can yield estimates with poor precision (Wang-Clow et al., 1995), and, to address this, additional structure might be specified for the relationship between the slope and intercepts and pattern. For example, one might assume the expected intercept is independent of pattern, and the expected slope is linearly related to the dropout time $t_k$ for pattern:

\[
b_{0i}^{(k)} = b_0, \quad b_{1i}^{(k)} = \nu_0 + \nu_1 t_k.
\]

This model is easily extended to include other covariates, such as indicators of treatment group, yielding formalizations of the conditional linear model of Wu and Bailey (1989).

**Model 8. An LOCF pattern-mixture model.** A common method for handling dropouts in longitudinal data, sometimes called “last observation carried forward” (LOCF) imputation, is to impute the missing values with last observation prior to dropout. This imputation method implements the idea that an individual’s outcome is unchanged after dropout, an assumption that needs to be checked for plausibility in real settings. Aside from the realism of the implied model for dropout, the LOCF method has the problem that if the outcome has some within-subject variation due to random fluctuations or measurement error, then imputing exactly the same value as that recorded just before dropout is not realistic. As a consequence, analysis of the data imputed by LOCF does not propagate imputation uncertainty, and hence does not yield valid inferences, even if the underlying model of no change after dropout is reasonable.

This problem can be addressed by formalizing the idea of LOCF as a pattern-mixture model, where individuals are stratified by pattern of dropout, and the individual mean outcome is assumed constant after dropout, but values after dropout can fluctuate around that mean. The key feature of the model is that each individual $i$ has an underlying profile of expected values $\mu_{ij}$, $j = 1, \ldots, n_i$ that would be observable in the absence of measurement error. If the individual drops out at some time $\tau_i$, then $\mu_{ij} = \mu_{i\tau_i}$ for all $j > \tau_i$; that is, the underlying mean remains unchanged after dropout. As a simple example of a model of this
kind, consider a homogeneous sample of \(N\) individuals with at most \(n\) repeated measures \(\{y_{ij}, j = 1, \ldots, n\}\). Let \(d_i\) be the number of measures for individual \(i\), and assume that

\[
(y_{ij} | \mu_{ij}, \sigma^2, d_i = d) \sim N(\mu_{ij}, \sigma^2)
\]

\[
\mu_{ij} = \begin{cases} 
\beta_0i + \beta_1ij, & \text{if } j < d, \\
\beta_0i + \beta_1id, & \text{if } j \geq d,
\end{cases}
\]

\[
(\beta_0i, \beta_1i | \nu_1)' \sim N((\beta_0i, \beta_1i)', \Gamma).
\]

This is an LOCF model with a linear profile up to the time of dropout. Again, extensions to include baseline covariates like treatment are readily formulated. This model could be used to multiply impute values of \(Y\) after dropout, yielding inferences that propagate imputation uncertainty.

18.7 Conclusion

I have reviewed likelihood-based methods for the analysis of models for longitudinal data with missing values. An important distinction is between models that ignore the missing-data mechanism, and hence assume the data are MAR, and models that relax the MAR assumption and incorporate assumptions about the missing-data mechanism. In many respects, ML and Bayes inferences for ignorable models are similar to corresponding inferences with complete data. The difference is that the likelihood is often more complicated, making computation more of a challenge, results are potentially more sensitive to model misspecification, and asymptotic results may be less valid because the log-likelihood function is not quadratic. Consequently, Bayesian inference based on the posterior distribution and relatively non-informative priors is attractive because it is less dependent on large sample sizes and deals in an appropriate way with nuisance parameters.

Non-ignorable models are more challenging because problems with lack of identifiability of the parameters are often severe, and assumptions about the mechanism leading to missing values need to be incorporated in the analysis. In selection models this requires an explicit parametric model for \(R\) given \(Y\); in certain pattern-mixture models the form of the model does not have to be explicit because assumptions about the mechanism are incorporated implicitly through restrictions on the parameters across patterns. Successful modeling requires realistic assumptions about the mechanism, which implies that information about the reasons why values are missing should be determined when possible and included in the analysis. For example, if some cases that drop out are plausibly MAR but others are not, it is better to build a model that reflects these different mechanisms than to assume the same MAR or NMAR mechanism for all dropouts. In general I think non-MAR situations are often best handled by relatively simple sensitivity analyses, where the assumptions are transparent. For example, if a subset of the dropouts are thought to have an NMAR mechanism, the model might assume the mean of the predictive distribution of those values deviates from the distribution assumed under MAR by some specified amount, say 0.2 or 0.5 times the residual standard deviation given known variables for that case. The results from “tilting” the MAR model in this way can then be assessed. Others (e.g., Horowitz and Manski, 2000) have advocated a sensitivity analysis over the full range of possible values of the missing values. This conservative approach is only feasible for missing variables that have a restricted range, such as binary or ordinal data, and the results are arguably too dispersed to be very useful unless there is a small number of missing values. A Bayesian analysis based on a subjective prior distribution relating distributions for non-respondents from distributions for respondents is in my view conceptually more satisfying (e.g., Rubin, 1977), although the challenge remains of incorporating reasonable departures from MAR in this prior specification.
The models considered here have assumed that missing data are confined to the repeated measures \( Y \), and covariate information \( X \) is fully observed. Current ML software does not allow for missing values in \( X \), so if values of covariates are missing then some additional work is needed to address that problem. One option is simply to drop values with missing covariates, which has advantages when the missingness mechanism depends on the values of the missing covariates themselves, but is wasteful of information and can result in bias if the mechanism is MAR. Another approach is to impute values of the missing covariates based on a joint model for the missing covariates given the observed covariates and \( Y \). Multiple imputation is recommended if this option is contemplated, so that imputation error is propagated. With relatively small numbers of missing covariates, a relatively simple model might suffice; as the fraction of missing values increases, more attention to specifying this model correctly is needed. The sequential multiple imputation (MI) methods discussed in Chapter 21 provide a useful tool for this multiple imputation step. Once the covariates are filled in, longitudinal models as discussed above can be fitted to the filled-in data sets, and results combined using MI combining rules discussed in Chapter 21.

Multiple imputation of missing values of \( Y \) under an explicit parametric model, as discussed in Chapter 21, is closely related to Bayesian inference based on the posterior distribution for that model. An advantage of MI is that the model for generating the multiple imputes may be different from the model used in the analysis; for example, in a clinical trial setting, the MI model may condition on information about side effects that are not part of the substantive models of interest, which focus on primary clinical trial outcomes.

References


Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5, 475–492.


REFERENCES

informative censoring: Modeling the number of observations as a geometric random variable.
Springer.
for repeated outcomes in the presence of missing data. *Journal of the American Statistical
Association* **90**, 106–121.
Shih, W. J., Quan, H., and Chang, M. N. (1994). Estimation of the mean when data contain
non-ignorable missing values from a random effects model. *Statistics and Probability Letters*
**19**, 249–257.
regression models for longitudinal binary response data with informative dropout. *Biometrics*
**54**, 367–383.
logistic regression models for multiple longitudinal binary functional limitation responses with
Wu, M. C. and Bailey, K. R. (1989). Estimation and comparison of changes in the presence of
Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of
informative right censoring by modeling the censoring process. *Biometrics* **44**, 175–188.