Incomplete Longitudinal Data Analysis with Multiple Partial Imputation

September 2005

BayesSoft Inc.
Los Angeles, California
Multiple Partial Imputation

Guide to Statistics

© BayesSoft Inc.

www.BayesSoft.com
3641 Midvale Ave. #207
Phone 310-600-4368

MPI 2.0 is a BayesSoft software package for incomplete longitudinal data analysis. This technical report serves as a companion to the User’s Guide for MPI 2.0.
# TABLE OF CONTENTS

INTRODUCTION ........................................................................................................ 1

CHAPTER 1
---- MISSINGNESS PATTERNS ........................................................................... 21

1.1 Intermittent Missingness vs. Dropout................................................................. 21
1.2 Missingness Patterns and Longitudinal Data Analysis ................................. 24
1.3 Overall Missingness Patterns Across Treatment Conditions .................... 25
1.4 Intermittent Missingness Patterns across Treatment Conditions ............... 30
1.5 Dropout Patterns across Treatment Conditions ............................................ 31

CHAPTER 2
---- MISSINGNESS MECHANISMS ................................................................. 34

2.1 Ignorable and Nonignorable Missingness Mechanisms .............................. 36
2.2 Missingness Mechanisms and Selection of Modeling Strategies ................ 37
2.3 Test of MCAR for Intermittent Missing Data ............................................... 40
2.4 Assessing Dropout Mechanisms within the Framework of MPI ................ 43

CHAPTER 3
---- MULTIPLE PARTIAL IMPUTATION .......................................................... 50

3.1 Taxonomy of Missing Data Methods ............................................................. 50
3.2 Multiple Imputation and Multiple Partial Imputation ................................. 53
3.3 Consolidating Multiple Analysis Results using Rubin’s Rule ................. 56
3.4 Other Computational Considerations in Applying MPI ............................ 57
3.5 Application of MPI to Data from the Smoking Cessation Study ............... 66

CHAPTER 4
---- SELECTION MODEL FOR NONIGNORABLE DROPOUT ...................... 73

4.1 Selection model and Pattern-Mixture ........................................................... 74
4.2 D-K models ................................................................................................. 76
4.3 Sensitivity Analysis Using D-K Models ...................................................... 78
4.4 Longitudinal Analysis for the Smoking Cessation Study ....................... 79
4.5 Longitudinal Analysis for the Milk Protein Trial Study ............................ 87
4.6 A Simulation Study .................................................................................... 91

CHAPTER 5.
---- PATTERN-MIXTURE MODEL FOR NONIGNORABLE DROPOUT ....... 94

5.1 P-M Models .............................................................................................. 94
5.2 P-M Models with Identified Parameters ................................................... 95
| 5.3 P-M Models with Identification of Restrictions | 96 |
| 5.4 Identification of Restrictions within the Framework of MPI | 98 |
| 5.5 Application of P-M to the Smoking Cessation Study | 100 |

CHAPTER 6.
--- RANDOM-EFFECTS MARKOV TRANSITION MODELS FOR NONIGNORABLE INTERMITTENT MISSINGNESS & DROPOUT

| 6.1 Incomplete Longitudinal Data Analysis | 109 |
| 6.2 REMTMs for Nonignorable Intermittent Missingness and Dropout | 113 |
| 6.3 MCMC Algorithms for Model Fitting and Imputation | 115 |
| 6.4 Application | 117 |

CHAPTER 7.
--- SUMMARY & DISCUSSION

REFERENCES | 125 |
INTRODUCTION

Substance abuse research is plagued by problems with missing data, especially in clinical trials of medical and behavioral therapies and in longitudinal surveys, where missed clinical visits and concomitant missing data points are common. Such behavior is part of the defining clinical features of the disorder, which, in turn, complicates measurement of treatment efficacy and even biases data analysis in testing the treatment efficacy. Recently, studies adopting longitudinal designs with a relatively large number of repeated measures have become common in substance abuse research.

A. LONGITUDINAL DATA ANALYSIS

A1. Longitudinal Data and Design

The defining characteristic of a longitudinal study is that each individual in the study is measured repeatedly through time. This is in contrast to cross-sectional studies where a single individual is measured only once.

In a typical longitudinal study, a maximum of $K$ repeated measures are collected on each of $N$ subjects (or participants in clinical trials). Let $Y_{ij}$ represent the response variable and $x_{ij}$ be the associated $P$-vector of explanatory variables observed at time $t_j$, for observation $j$ $(j = 1, ..., K)$ on subject $i$ $(i = 1, ..., N)$. The set of repeated outcomes for each subject $i$ are collected into a $K$-vector, $Y_i = (Y_{ij_1}, ..., Y_{ij_K})^T$, with mean $E(Y_i) = \mu_i = \mu(\beta)$ and covariance matrix $Var(Y_i) = \Sigma_i = \Sigma(\alpha)$. Similarly, $X_i = (x_{ij_1}, ..., x_{ij_K})^T$ denotes the design matrix for subject $i$. Here, we assume the distribution of the $Y_{ij}$’s, possibly after suitable transformation, is Gaussian, i.e.

$$Y_i \sim N(\mu_i, \Sigma_i).$$

Most longitudinal data analyses involve the linear regression model written in marginal form as

$$Y_{ij} = x_{ij}\beta + \epsilon_{ij}^*$$

where $\epsilon_{ij}^* \sim N(0, \sigma^2_{ij})$ are correlated heterogeneous error terms, $\beta = (\beta_1, ..., \beta_p)$ is a $P$-vector of unknown regression coefficients (e.g. treatment effects in clinical trials).
Approaches to longitudinal data analysis are influenced by the correlation structure among repeated measures observed for each subject. That is, the $\varepsilon_i$’s (or $Y_i$’s) are not assumed to be identically independently distributed (i.i.d.), as in the classical linear regression model. In the following discussion, we assume there are no correlations across individuals, or subjects. There are two main methods to analyze longitudinal data: time-naïve strategies (Hall et al. 2001) and longitudinal strategies (Diggle, Liang, and Zeger 1994).

A2. Time-naïve Strategies for Longitudinal Data Analysis

Time-naïve models, also known as “aggregation models,” (Yang and Shoptaw 2005) or “two-step methods” (Diggle, Liang, and Zeger 1994), collapse the repeated measurements into a scalar composite score and then traditional ANOVA or t-tests are performed. For example, as a partial solution to imbalance in the data (caused by missingness) and within-subject correlation, Ling and his colleagues (Ling et al. 1997) suggest the use of “treatment effectiveness score” (TES) and “joint statistical probability” (JSP) in urine data analysis. In applying TES, the number of urine data points indicating no recent drug use is counted for each person. Similarly in applying JSP, the proportion of such “clean” urine data points is calculated for each subject. In the field of substance abuse research, these methods were the only statistical solutions for longitudinal designs before the advent of advanced longitudinal modeling strategies, and more importantly, the availability of software packages to implement them.

The advantage of time-naïve methods is that they are simple and easy to interpret, but the time trend is lost when collapsing or aggregating the repeated-measure vector into a scalar. Another limitation for this modeling strategy is that it requires very strong missing-data assumptions.

A3. Longitudinal Strategies for Longitudinal Data Analysis

Modern approaches for longitudinal models, which are extensions of generalized linear models (GLMs: McCullagh and Nelder 1989), can be roughly categorized into three groups: marginal models, random effects models (also known as mixed models), and transition models. In each of these models, both the dependence of the response variables on the explanatory variables and the correlation among the response variables are modeled.

Marginal Models

Marginal models (Liang and Zeger 1986, Prentice 1988) are appropriate when inference about the population average is of interest. In substance abuse clinical trials, for example, investigators are usually interested in the average difference between control condition and treatment condition. In a marginal model the
marginal expectation of the response, \( E(Y_i) = \mu_i \), is modeled by the explanatory variables, \( x_{ij} \), as

\[
h(\mu_i) = x_{ij}\beta
\]

where \( h \) is a known link function. The correlation or variance-covariance among repeated observations, \( Y_{i1}, \ldots, Y_{ik} \), is modeled separately. For example, a compound symmetric correlation structure assumes that

\[
corr(Y_i, Y_a) = \begin{cases} 1 & \text{if } j = k \\ \alpha & \text{if } j \neq k \end{cases}
\]

where \( |\alpha| \leq 1.0 \) is viewed as a nuisance parameter. An algorithm for parameter estimation is suggested by the generalized estimating equation (GEE)

\[
\sum_{i=1}^{N} (\frac{\partial \mu_i}{\partial \beta})^T \text{var}(Y_i)^{-1} (Y_i - \mu_i) = 0.
\]

An attractive property of marginal models is that inferences about \( \beta \) are sometimes valid even though an incorrect form of covariance structure is assumed. For non-normally distributed (e.g. binary, categorical, or Poisson) repeated measures, marginal models using GEE are a popular option.

**Random Effects Models**

In real-life data sets, there is usually a natural heterogeneity between individuals. When making inferences, such individual-varying factors should be considered. Random effects models (Laird and Ware 1982, Breslow and Clayton 1993) are appropriate for such a purpose.

Let \( \gamma_i \sim N(0, \mathbf{G}) \) denote the unobserved random effects for the \( i^{th} \) individual, which are assumed independent of each other. Then the original linear model can be rewritten in the form

\[
Y_i = x_{ij}\beta + d_{ij}\gamma_i + \varepsilon_i
\]

where \( \varepsilon_i \sim N(0, \sigma^2) \) represent i.i.d. residual errors and the covariates included in \( d_{ij} \) are usually a subset of the explanatory variables in \( x_{ij} \), that is, \( d_{ij} \subseteq x_{ij} \). For example, in a clinical trial, a random intercept and slope model is a special type of random effects model, where \( d_{ij} = (1, t_{ij}) \) and \( \gamma_i = (\gamma_{i1}, \gamma_{i2})^T \) where \( \gamma_{i1} \) is the unobserved random intercept effect and \( \gamma_{i2} \) is the unobserved random slope effect.

In random effects models, the correlation between repeated measures is explained by the random effects (i.e., \( \gamma_i \)). Expressed in vector-matrix form, the unconditional variance of \( Y_i \) is

\[
Var(Y_i) = d_i \mathbf{G} d_i^T + \sigma^2 \mathbf{I},
\]
and the joint density function of observations (i.e., \( Y_i \)) and random effects (i.e., \( \gamma_i \)), conditioned on the unknown but fixed parameters \( \beta, \sigma^2 \), and \( G \), is given by

\[
f(Y_i, \gamma_i | \beta, \sigma^2, G) = f(Y_i | \gamma_i, \beta, \sigma^2, G)f(\gamma_i | \beta, G)
\]  

\[
= \prod_{j=1}^{K} f(Y_{ij} | \gamma_i, \beta, \sigma^2)f(\gamma_i | G)
\]

(8)

since \( Y_{i1}, \ldots, Y_{iK} \) are conditionally independent given random effects \( \gamma_i \). The likelihood function for \( \beta, \sigma^2 \), and \( G \) is

\[
L(\beta, \sigma^2, G | Y) = \prod_{i=1}^{N} f(Y_i | \beta, \sigma^2, G)
\]

\[
= \prod_{i=1}^{N} \int f(Y_i, \gamma_i | \beta, \sigma^2, G) d\gamma_i
\]

\[
= \prod_{i=1}^{N} \prod_{j=1}^{K} f(Y_{ij} | \gamma_i, \beta, \sigma^2)f(\gamma_i | G)d\gamma_i
\]

(9)

which is a marginal distribution of \( Y \) after integrating out the random effects from the joint distribution. Applying numerical optimization methods to this likelihood function, maximum likelihood estimates can be obtained. Usually this is difficult to do because the integration require numerical methods and there is usually no analytical derivatives in the optimization procedure. MCMC-based Bayesian inference provides an appealing alternative solution, where the random effects can be viewed as missing values and conveniently simulated.

A final comment for random effects models is on their relationship with mixed models, which have similar forms to random effects models, that is,

\[
Y_i = X_i \beta + Z_i \gamma_i + \varepsilon_i,
\]

(10)

where \( Z_i \) is a known design matrix corresponding to the random effects. In a narrow sense, mixed models are defined as models that classify the effects of design factors by a mixture of fixed effects and random effects. In this narrow sense, we sometimes do not make a distinction between the two types of models: mixed models and random effects models. Recently, the meaning of “mixed models” has been extended to a wider range to include other types of models such as covariance pattern models and random coefficient models, although such terminology is not rigorous. More accurately, the random effects models defined by Laird and Ware (1982) and shown above can be viewed as a special type of mixed model.

In our discussion within the scope of this technical report for Gaussian distributed repeated measures, we treat the two terms to be interchangeable although “mixed models” tends to be more general and broader. Such an interchangeable view reflects the situation that in many practical settings “random effects model” and “mixed model” are considered to be equivalent. For more details on the historical developments of the mixed model, refer to Henderson and Serle (1990), McCulloch and Searle (2001), and Brown and Prescott (1999).
Transition Models

In a transition model (Korn and Whittermore 1979, Cook 1999, Diggle, Liang, and Zeger 1994), the within-subject correlation is modeled by making use of the past responses to predict the present and future ones. To simplify notation, let \( H_j = \{Y_{i1}, ..., Y_{i(j-1)}\} \) represent the preceding \( j-1 \) observations or the “history” for \( Y_i \). The transition model can be specified as a regression model in the form of

\[
\begin{align*}
    h(\mu_y | H_i) &= x_i^T \beta + f_H(H_i, \alpha^*) \\
    \text{where } f_H(H_i, \alpha^*) &\text{ is a function of the history with parameters } \alpha^*.
\end{align*}
\]

(11)

It turns out that the correlation between any two within-subject repeated measures, \( Y_{i1} \) and \( Y_{ik} \), can be difficult to calculate, although the above form of the models seems quite simple.

The likelihood function for transition models is equal to the joint probability function,

\[
L(\alpha^*, \beta | Y) = \prod_{i=1}^{N} f(Y_{i1} | \alpha^*, \beta) f(Y_{i2} | Y_{i1}, \alpha^*, \beta) \cdots f(Y_{ik} | Y_{i1}, ..., Y_{i(k-1)}, \alpha^*, \beta).
\]

(13)

Note that for a history function involving \( S \) past values, the first \( S \) terms of this likelihood need to be handled in a special manner as there are not \( S \) preceding values until we reach the term for \( Y_{i(S+1)} \).

Transition models have a dynamic aspect and the local transition probabilities can be obtained to make local inferences. In practice, for binary or categorical longitudinal data, transition models, jointly with marginal models, provide powerful tools. For continuous normal-distributed models, transition models can be used as a supplemental tool for random-effects models.

A4. Software for Longitudinal Models

For normal data, random-effects models can be fitted by the MIXED procedure in SAS, lme in Splus, or XTREG in STATA. For non-normal data, however, there are still no completely satisfactory packages, although the NLINMIX procedure and the GLIMMIX macro in SAS and nlme in Splus can be used to model some generalized linear mixed models (GLMM). For marginal models using GEE, GENMOD in SAS, Oswald in Splus, and XTGEE in STATA can be applied to data with exponential family distributions. For standard transition models, most software packages with linear model functionality can be used by moving the preceding observations into the design matrix.
B. MISSING DATA PROBLEMS IN LONGITUDINAL STUDIES

The missing data problem is especially critical in longitudinal data analysis since the structure of missingness is complicated, involving both intermittent missingness and dropout. Intermittent missingness happens when participants occasionally miss their clinical visits but they still remain in the study. Dropout occurs when participants prematurely withdraw from the study.

Missing data require special treatment prior to analyzing for intervention effects. In practice, however, many data analysts select analysis strategies without considering or testing assumptions about the missingness mechanisms. Identification of these assumptions about the missing data are critical in order to obtain unbiased parameter estimates. In this report, missing data problems in repeated-measures studies are fully discussed, with special attention paid on dropout issues. By summarizing theoretical solutions to the general missingness and dropout problems, we hope to provide a systematic framework to handle missing data problems in repeated-measure substance abuse studies. A data set from a smoking cessation study will be used for illustration purposes. Most of the analysis in this report was performed using our own software package MPI 2.0 (see User’s Guide, BayesSoft Inc. 2005); some of the analysis was conducted using other packages especially SAS and Splus/R.

B1. Missingness Patterns in Longitudinal Study

When we study missing data problems, we often mention two terms: missingness pattern and missingness mechanism. The pattern tells you which entries in the data matrix were unavailable, and the mechanism describes why they were missing.

1). General Pattern

<table>
<thead>
<tr>
<th>?</th>
<th>?</th>
<th>?</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

2). Monotone Pattern

<table>
<thead>
<tr>
<th>?</th>
<th>?</th>
<th>?</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

Figure 1. Missingness patterns ("?" represents a missing item in the data matrix)

When some repeated measures in \( Y = (Y_y) \) are not observed, the data matrix \( Y \) can be partitioned into \( Y = (Y^{(o)}, Y^{(m)}) \), where \( Y^{(o)} \) denotes the data observed or collected while \( Y^{(m)} \) represents the data that would be observed if they were not missing. \( Y^{(m)} \) is, of course, a theoretical construct, since we do not observe these
data points! A missing-data indicator matrix, $\mathbf{R} = (r_{ij})$, can be used to represent the missingness patterns of $\mathbf{Y}$, where element $r_{ij} = 1$ or 0 corresponding to $Y_i$ observed or missed, $(i = 1, \ldots, N; j = 1, \ldots, K)$.

There are two types of missingness patterns, general patterns and monotone patterns, which are shown in Figure 1. In an incomplete longitudinal data set, intermittent missing data usually associate with general patterns, and dropouts always display monotone patterns after being sorted. Monotone patterns are separated from general patterns because this special pattern offers a computational advantage: the likelihood for the whole data can be factored into a product of chain-like series of conditional likelihoods, so that faster algorithms exist in performing multiple partial imputation (MPI), which will be introduced later in this report.

**Patterns for Dropouts**

For any subject, the pattern of dropouts (i.e., missing data due to dropout) can be simply denoted by an integer $d_i = j$, $(j = 2, \ldots, K)$ indicating that the $i^{th}$ subject in a study prematurely withdrew at time point $t_i$. For a subject that did not withdraw, we set $d_i = K + 1$.

**Patterns for Intermittent Missing Data**

Patterns for intermittent missing data can be denoted as above by a 0/1 vector $\mathbf{R}_i = (r_{i1}, \ldots, r_{Id_i-1})$, where $r_{ij} = 1$ or 0 corresponding to $Y_i$ observed or missing. The length of $\mathbf{R}_i$ varies across subjects depending how long they remain in the study before withdrawal.

**B2. Missingness Pattern and Longitudinal Data Analysis**

Missingness pattern plays a critical role in longitudinal data analysis especially when analysis depends on various time naïve methods. As we mentioned earlier, time-naïve models require very strong assumptions for missing data. For example, when TES (Ling et al. 1997) is used in analyzing a binary longitudinal data set, the number of “clean” urine data points (by “clean”, we mean “free of drug metabolite”) is summarized for each subject, and then these numbers for all the participants are used to perform a t-test between treatment and control groups. If the participants in the treatment group have a higher rate of missing data, then the testing on treatment efficacy would be conservative. However, in many studies, participants in the control group may have a higher missingness rate and the testing would lead to the unintended consequence of favoring the treatment condition over the control condition. By summarizing the number of “clean” urines in applying TES, we implicitly adopt the imputation of the missed urine data points as “dirty”. Such a “missingness as dirty” imputation is not recommended unless you have strong reasons to do so.
For advanced longitudinal modeling strategies, we should also pay attention to the distribution of the missingness patterns across treatment conditions, although an unbalanced or uneven distribution does not bias our analysis as much as time-naïve analysis. For example, in applying a random effects model to a treatment-control trial lasting 12 months and assuming that all participants in one group withdrew at the middle point and all participants in the other group completed the trial, we still can fit the model by using all the observed data. However, when making inferences, we should recognize the limit that the intervention effects cannot be tested for the second period of the trial unless certain assumptions are employed.

**Retention Analysis and Dropout Patterns**

As seen in many publications in the field of addiction medicine, *retention analysis* is popularly used when reporting clinical trial results. Retention analysis is equivalent to *survival analysis* for biostatisticians or data analysts in other medical fields where the lengths of participants’ stay (or “survival” lengths) in the study are compared across treatment conditions using statistical tests, such as the log-rank test (Hosmer and Lemeshow 1999). From a purely statistical perspective, such a comparison is only part of the whole analysis process and reflects only dropout patterns across treatment groups. Another term called “attrition,” as a complement to “retention,” is also popularly mentioned in practice by investigators and data analysts.

Here, we see how important it is to assess dropout patterns, since many investigators believe that longer “survival” lengths represent a positive treatment effect. For this report, we use the term “dropout”, although “attrition” or “retention” may be more familiar to some investigators.

**Modeling Intermittent Missingness Patterns**

When missingness status is viewed as a binary response (1=observed, 0=missing), we can apply almost any of the longitudinal models introduced earlier to test whether the patterns for intermittent missing data are evenly distributed across treatment groups. Specifically, for mixed models, generalized linear mixed models (GLMMs: McCulloch and Searle 2001) should be used. Such a test can be performed in the following way: (1) in the missingness indicator matrices (i.e., \( \mathbf{R} = (r_{ij}) \)), recode the dropouts (i.e., missingness due to dropout) as “NA”, (2) apply any longitudinal GLMM to these matrices and ignore the “NA”s. This procedure, jointly with retention analysis, is recommended especially when one plans to apply time-naïve modeling strategies.

**B3. Missingness Mechanisms**
A missingness mechanism, sometimes called missing-data process, describes the process or mechanism underlying the missingness. In practice, there are many factors that introduce missing data, e.g., traffic congestion, weather conditions, appointment conflicts, faulty data entry or collection, side-effects of the medicines, loss of interest in the intervention, etc. Theoretically, we can define missingness as the distribution of missingness indicators conditionally on the data themselves and a set of parameters, that is,

\[
\text{Missingness Mechanism} = P(R_i \mid Y_i, \psi).
\]  

For example, if the repeated measures are systolic blood pressure levels and a censoring value at 140 mmHg will cause a missing value for the \(i\)th subject, then we have \(\psi = 140\) mmHg and

\[
P(r_{ij} \mid Y_{ij}, \psi) = \begin{cases} 1 & \text{if } Y_{ij} < 140 \text{mmHg} \\ 0 & \text{Otherwise} \end{cases}
\]

In the context of likelihood-based inference and based on the terminology of Little and Rubin (1987), Yang and Shoptaw (2005) classify intermittent missingness mechanisms into the following four types.

1. **missing completely at random (MCAR):**

\[
P(R_i \mid X_i, Y_i, \psi) = P(R_i \mid \psi),
\]

i.e., “\(R_i\) is independent of covariates \((X_i)\), observed outcomes \((Y_i^{(o)})\), and missing outcomes \((Y_i^{(m)})\).”

2. **covariates-dependent missing at random (CMAR):**

\[
p(R_i \mid X_i, Y_i, \psi) = P(R_i \mid X_i, \psi),
\]

i.e., “\(R_i\) depends only on covariates \((X_i)\).”

3. **outcomes-dependent missing at random (OMAR):**

\[
p(R_i \mid X_i, Y_i, \psi) = P(R_i \mid X_i, Y_i^{(o)}, \psi),
\]

i.e., \(R_i\) depends on both covariates \((X_i)\) and observed outcomes \((Y_i^{(o)})\),” including a special case where \(R_i\) depends only on observed outcomes \((Y_i^{(o)})\).

4. **Informative missingness (IM) or nonignorable missingness (NM)**

\[
p(R_i \mid X_i, Y_i, \psi) = P(R_i \mid Y_i^{(o)}, \psi)
\]

i.e., “\(R_i\) depends at least on the unobserved values of the outcomes \((Y_i^{(m)})\).”

For dropouts, the mechanisms have a classification similar to that for intermittent missing data, except that in describing the dropout process, we assume that dropouts are dependent only on combinations of the following three components: covariates, previous observed repeated outcomes, and measured outcomes at dropout. It is rarely meaningful to let the dropouts depend on the future observations, especially in clinical trials. Even if they did, all the future outcomes are missing data and usually do not contribute to making inferences on the parameters of interest. Assuming either there are no intermittent missing data or
they have been imputed (the case in applying multiple partial imputation), we classify the dropout processes into:

(1) dropout completely at random (DCAR),

(2) covariates-dependent dropout at random (CDAR),

(3) preceding-outcomes-dependent dropout at random (PDAR),

(4) informative dropout (ID),

which are parallel to MCAR, CMAR, OMAR, and IM, respectively, with similar definitions. Such a specific classification aims to emphasize the difference between intermittent missing data and dropouts. Table 1 summarizes the mechanisms for missing data in longitudinal studies.

<table>
<thead>
<tr>
<th>For Intermittent Missing Data</th>
<th>For Dropouts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ignorable</strong></td>
<td></td>
</tr>
<tr>
<td>MCAR</td>
<td>DCAR</td>
</tr>
<tr>
<td>Missing Completely At Random</td>
<td>Dropout Completely At Random</td>
</tr>
<tr>
<td>(P(R_i \mid X_i, Y_i, \psi) = P(R_i \mid \psi))</td>
<td>(P(d_i \mid X_i, Y_i, \psi) = P(d_i \mid \psi))</td>
</tr>
<tr>
<td>CMAR</td>
<td>CDAR</td>
</tr>
<tr>
<td>Covariates-dependent Missing At Random</td>
<td>Covariates-dependent Dropout At Random</td>
</tr>
<tr>
<td>(p(R_i \mid X_i, Y_i, \psi) = P(R_i \mid X_i, \psi))</td>
<td>(p(d_i \mid X_i, Y_i, \psi) = P(d_i \mid X_i, \psi))</td>
</tr>
<tr>
<td>OMAR</td>
<td>PDAR</td>
</tr>
<tr>
<td>Outcomes-dependent Missing At Random</td>
<td>Preceding-outcomes-dependent Dropout At Random</td>
</tr>
<tr>
<td>(p(R_i \mid X_i, Y_i, \psi) = P(R_i \mid X_i, Y_i^{(\psi)} \psi))</td>
<td>(p(d_i \mid X_i, Y_i, \psi) = P(d_i \mid X_i, H_{id_i}, \psi))</td>
</tr>
<tr>
<td><strong>Non-ignorable</strong></td>
<td><strong>ID</strong></td>
</tr>
<tr>
<td>IM</td>
<td>Informative Dropout</td>
</tr>
<tr>
<td>Informative Missing</td>
<td>Informative Dropout</td>
</tr>
<tr>
<td>(p(R_i \mid X_i, Y_i, \psi) = P(R_i \mid X_i, Y_i^{(\psi)} \psi))</td>
<td>(p(d_i \mid X_i, Y_i, \psi) = P(d_i \mid X_i, H_{id_i}, Y_{id_i}, \psi))</td>
</tr>
</tbody>
</table>

Note: For subject \(i\), who drops out of a study at \(t_{id_i} \) (\(d_i \in \{2,3,...,K\}\)), \(H_{id_i} = (Y_{id_i},...,Y_{id_i}), Y_{id_i}\) denotes a vector of observed preceding outcomes, \(Y_{id_i}\) denotes the value that would be observed at \(t_{id_i}\) if the subject did not drop out at that time.

**Table 1.** Missingness mechanisms for data from longitudinal studies

In the above discussion, we stated that dropouts rarely depend on future observations, but we do in certain situations observe such a relationship which Little (1995) calls “nonignorable random-Coefficient-Based Dropout”. Such a type of nonignorable dropout mechanisms can be viewed as another form of informative dropout (ID), which will be exclusively addressed in Chapter 6. The ID defined in Table 1 actually was named “nonignorable outcome-based dropout” by Little (1995).
Properties of methods of handling missing data are strongly influenced by assumptions made about the missingness mechanism. It is important to consider the likely nature of this mechanism in a particular application.

**Full Likelihood Function**

In order to obtain valid or unbiased inference from an incomplete longitudinal data set, we should model both the observed data and missing data in a joint model. For subject $i$, such a joint model can be represented by $p(Y_i, R_i | X_i, \theta, \psi)$, where, as above, $\psi$ and $R_i = (r_{i1}, ..., r_{ik})$ define the missingness mechanism and pattern, and $\theta = (\beta, \alpha)$ is the set of parameters for the longitudinal model. The full form of the likelihood function is

$$L_F(\theta, \psi | Y^{(o)}, R) \propto \prod_{i=1}^{N} \int p(Y_i^{(o)}, R_i, Y_i^{(m)} | X_i, \theta, \psi) dY_i^{(m)}. \quad (19)$$

Most likelihood-based longitudinal models in popular general-purpose software packages, such as SAS, STATA, and SPSS, base their inference on a reduced form of the likelihood function

$$L_R(\theta | Y^{(o)}) \propto \prod_{i=1}^{N} \int p(Y_i^{(o)}, Y_i^{(m)} | X_i, \theta) dY_i^{(m)}. \quad (20)$$

Inferences based on the reduced form of the likelihood are valid, however, only when the missingness mechanisms are ignorable (e.g., OMAR for intermittent missing values and PDAR for dropouts). When we say that a type of missingness mechanism is “ignorable,” we mean that the full joint likelihood function for $\psi$ and $\theta$ can be expressed as a product of two functions, one for each. In this case, the inference for $\theta$ can be made from $L_R(\theta | Y^{(o)})$ while ignoring $\psi$. For example, in SAS (proc MIXED) and Splus (lme), random-effects models require that the missing data are ignorable. For more information on ignorable missingness, refer to Rubin (1976), Little and Rubin (1987), Schafer (1997), and Verbeke and Molenberghs (2000).

As many studies suggest (Little and Rubin 1987, Belin et al. 2000, Molenberghs and Verbeke 2001, and Schafer 1997), when the number of variables in the data set is large, the ignorability assumption is approximately valid. This is common in analyzing survey data, where non-monotoneic patterns of missing values are more common than dropouts. However, in incomplete substance abuse data sets, dropouts are more common, and are usually nonignorable. When applying random effects models or other types of mixed models without further assessment on the dropout mechanism, biased estimation may result (Diggle and Kenward 1994, Yang and Shoptaw 2005, Hall et al. 2001).

When analyzing data through other modeling strategies such as marginal models with quasi-likelihood estimation and Markov Chain transition models, missing
data cause more difficulties. For example, marginal models using GEE are sensitive to missing data, requiring stringent assumptions on the missingness mechanisms (Paik 1997). Transition models based on Markov processes may be more robust to missing data, but they have a limitation. In modeling the transition probability among consecutive data points, a data point is considered only when either the data point ahead of it or the one following it is observed. When both are missing, the “present” data point is ignored in fitting the model, although itself is observed (Yang et al 2004).

More hidden problems exist for inference based on time-naïve methods, since the missingness mechanisms become seriously distorted when the repeated high dimension data points are compressed into composite scores. Conservatively, we anticipate that MCAR or DCAR is required in applying these aggregation methods (Yang and Shoptaw 2005).

B5. Longitudinal Models Handling ‘Nonignorable’ Missingness

As discussed above, in order to handle nonignorable missingness, we should theoretically base our analysis on the full form of the likelihood function. The joint model \( p(Y_i, R_i | X_i, \theta, \psi) \) can be conveniently factored into two forms:

\[
(A) = p(Y_i | X_i, \theta) p(R_i | X_i, Y_i, \psi)
\]

and

\[
(B) = p(Y_i | R_i, X_i, \theta) p(R_i | X_i, \psi). \tag{22}
\]

Accordingly, there exist two theoretical possibilities for conducting a valid or unbiased analysis, which do not ignore the missingness mechanisms. The first is to analyze models in format A, which are called selection models (Diggle and Kenward 1994, Little 1995, Molenberghs and Verbeke 2001, Li et al. 2005). The second is to analyze models in format B, which are called pattern-mixture models (Little, 1993, Little 1994, Yang et al. 2005). Missingness mechanisms have clearer definitions in selection models than in pattern-mixture models.

Selection Models

Much of the early development of selection models originates from the econometrics literature. In 1976, Heckman proposed the tobit model, which combines a marginal Gaussian regression model for the response with a Gaussian-based threshold model for the probability of missing a value. In 1994, Diggle and Kenward developed a selection model for longitudinal data with dropouts (hereafter the D-K model), which is actually a variant of the tobit model. By combing the multivariate Gaussian linear model with a logistic dropout model, the D-K model can be used for full-likelihood based analysis. For optimizing the full likelihood function, Diggle and Kenward (1994) employed the Nelder and Mead simplex algorithm (Nelder and Mead 1965), but this algorithm converges very slowly especially when the number of repeated measures is large with a high proportion of missing data. The D-K model can also be viewed as a transition model, which is capable of handling outcome-based nonignorable dropout (Little
The D-K model has been implemented in an Splus library called Oswald but its unstable performance prohibits it from routine usage in substance abuse data analysis. In Chapter 4, we discuss the D-K model further.

**Pattern-Mixture Models**

Recently, the high sensitivity of selection model-based inference to the correct specification of the distribution of the repeated measures as well as the mechanisms of dropout has been extensively documented. This has led to a growing interest in pattern-mixture modeling (P-M model hereafter); see Little (1993, 1994, 1995), Hogan and Laird (1997), Ekholm and Skinner (1998), Molenberghs, Michiels, and Kenward (1998), etc. Suppose in a study with at most three time points, a Gaussian-distributed response variable is measured for each subject and dropouts are observed on the second and third time points. In applying P-M models, we can stratify the subjects into three groups according to the three missingness patterns as shown in Figure 2. Within each group, the mean vector and covariance matrix are modeled using random-effects models,

\[ Y_j | d_j, X_j \sim N(\mu(d_j, X_j), \Sigma(d_j, X_j)) \]

and the dropout mechanism is modeled using

\[ p(d) = \pi_d = f(d_j | X_j, \psi) \]

where \( d_j = 2, 3, 4 \) indicating dropout time at time 2, 3, and 4 (i.e., no dropout). Then the marginal distribution of the response is a mixture of normal distributions with mean

\[ \mu = \sum_{d=2}^{4} \pi_d \mu(d) \]

and a variance matrix that can be derived by applying the delta method along with certain restrictions on the covariance parameters associated with dropouts. For more details, refer to Little (1993).

<table>
<thead>
<tr>
<th></th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern-1</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>Pattern-2</td>
<td>O</td>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td>Pattern-3</td>
<td>O</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Note: O=being observed; X=being missing.

**Figure 2.** Missingness patterns in a 3-measure data set with dropouts

In substance abuse studies with longitudinal designs, the number of repeated measures is usually large (e.g. >10) as is the number of missingness patterns (involving both intermittent missingness and dropout). A P-M model in its original sense is not practical since the number of subjects in each missingness pattern would be too small. However, by applying the strategy called multiple partial imputation (MPI), we may impute the intermittent missing values from the data sets allowing for the fitting of a P-M model. In certain cases, we can further decrease the number of dropout patterns through a data subset selection procedure that is similar to a bootstrap approach. In Chapter 5, a two-stage version of MPI
will be introduced where dropouts are further imputed after the partial imputations are conducted for intermittent missing values.

C. MULTIPLE PARTIAL IMPUTATION

A popular incomplete data analysis strategy is to use imputation techniques to replace the missing data with plausible values, thus rendering the data set complete and allowing for complete data analysis. An advantage of the imputation method is that the imputation procedure can be separated from the analysis procedure so that imputers can take advantage of available auxiliary information or confidential information that is blocked to analysts.

C1. Multiple Imputation (MI)

A criticism for single imputation methods is that by filling the missingness space-holders with some concrete values and treating them as observed data, the variances in parameter estimation would be underestimated. In order to reflect such uncertainty introduced by missing data, Rubin (1987) proposed multiple imputation (MI), a Monte Carlo technique, in which the missing data are replaced by $m > 1$ sets of simulated values. Under MI, $m$ complete data sets are thus created and each of them is analyzed by standard complete-data methods such as a random effects model. Then, the results are combined using a special rule to produce estimates and confidence intervals that incorporate missing-data uncertainty (Rubin 1987). Schafer (1997) has developed a series of algorithms based on Monte Carlo Markov Chain (MCMC) methods to create multiple imputations for multivariate continuous data, categorical data, or mixed data where both types exist.

C2. Multiple Partial Imputation (MPI)

From the above discussion on incomplete longitudinal data analysis, we see that both patterns and mechanisms of missing data are critical in selecting and fitting longitudinal models. When longitudinal missing data are further divided into intermittent missing data and dropouts according to their visual patterns, we have more careful definitions of their mechanisms. Based on our statistical analysis experience in substance abuse research, dropouts usually occupy a larger proportion of overall missingness and there are usually study-related reasons behind their missingness. In other words, for participants receiving active treatment or intervention, dropouts are rarely completely at random (i.e. DCAR). Instead, they are quite likely to be informative (i.e. ID). Also, considering that intermittent missing values are usually sparsely scattered within the data set, even if they are not MAR or if they are "nonignorable", we may still have confidence...
in imputating them, especially when the number of repeated measures is large (Belin et al. 2000).

Theoretically, there are two ways to apply MI for incomplete longitudinal data sets with nonignorable or informative dropouts. One way is to develop imputation algorithms that take into consideration the dropout mechanisms. This approach works well when the mechanism is clear to the imputers. For example, if the investigators in a study are sure that people withdraw because their systolic blood pressure is over a censoring level (e.g. 180 mmHg). Unfortunately, this type of analysis can be performed only in situations such as laboratory experiments using animals or cohort studies with participants that can be followed easily after their withdrawal. This case is specifically discussed in Chapter 5.

The other way is to make multiple imputations only for intermittent missing data but leave dropouts as they are. Then the multiple partially-imputed sets of data are analyzed using longitudinal models such as D-K or P-M models that can handle nonignorable dropouts. We call this strategy *Multiple Partial Imputation* (MPI; Yang and Shoptaw 2005) and have implemented it into our software package named MPI 2.0. By applying MPI, the dropouts are isolated from intermittent missing data and can be further investigated and modeled. As we mentioned earlier when applying selection or pattern-mixture models, the number of missingness patterns cannot be too large, or else we would not have enough subjects within each pattern. By partially imputing the intermittent missing data, we can significantly decrease the number of missingness patterns. For example, in a longitudinal study with 8 repeated measures on each subject, we can have as many as \(256 = 2^8\) patterns of missingness. After performing partial imputation for intermittent missing data, the maximum number of missingness patterns becomes 8, which is the possible number of dropout patterns. Therefore, MPI provides a computational solution for applying D-K or P-M models. In the smoking cessation study considered in Chapter 4, when the D-K models were fitted without partial imputation, we observed convergence problems. Within the framework of MPI, however, convergence was achieved.

Another advantage of MPI is that the dropouts can be separated from intermittent missing data and their dropout mechanism can be investigated. After partial imputation of intermittent missing data, we can apply testing procedures to the partially imputed data sets to test the assumption of dropout mechanism. In Chapter 3, a Monte Carlo Markov Chain algorithm for creating multiple partial imputations is described along with some discussion on convergence diagnosis. In the same chapter, we also introduce Rubin’s rule to consolidate multiple analyses and make MPI-inferences.

**C3. Assessing Missingness Mechanisms and Selecting Appropriate Modeling Strategies**

In choosing a proper modeling strategy for incomplete longitudinal data, it is essential to do some assessment on the missingness mechanisms since different
modeling strategies require different assumptions. For all possible mechanisms listed in Table 1, a complete investigation should involve both statistical testing and practical investigation such as a follow-up survey in the national census for undercount estimation (Belin et al. 1993). In substance abuse clinical trials, however, such missing-data related follow-up surveys are seldom performed since they are either resource demanding or not ethical. Fortunately, in many studies, we can partially assess missingness mechanisms by statistical testing applied to the collected data and the missingness pattern matrix, which jointly may provide evidence for rejecting certain missing-data assumptions. In this report and MPI 2.0, we focus on such statistical testing methods to assess missingness mechanisms. An important comment here is that, in any of these statistical tests, any significant testing result (e.g., P-value<0.05) provides strong evidence in rejecting an assumption (e.g. MCAR or DCAR), but an insignificant result (e.g., P-value > 0.05) does not mean that we should accept the assumption. This is not only due to the definition of p-values but also because the test is based only on the observed data. When the proportion of missingness is large or when the sample size is small in a longitudinal data set, the observed data themselves may not provide enough information on the missingness mechanism. Therefore, we should always adopt a conservative perspective in explaining these testing results.

Tests of Missingness Mechanisms for Intermittent Missing Data

For intermittent missing data, the first task is to test whether the missingness mechanism is CMAR (including the case of MCAR), which is the assumption in applying marginal models such as those using GEEs. Under the assumptions of MCAR or CMAR, most other longitudinal models techniques and time-naïve methods can be applied. Even certain ad-hoc missing-data methods, such as regression mean imputation or hot-decking, can be applied under these mechanisms.

When testing the mechanisms for intermittent missing data, let us temporarily ignore the dropouts by excluding from consideration those participants who have dropouts. Such an exclusion procedure can be performed in two ways: the omnibus test or the pseudo bootstrap test. In the omnibus test, a statistical test is performed on the original data set by including only completers (i.e., participants who have only intermittent missing data). In a pseudo bootstrap test, a large number of random selected subsets are assessed to test the missingness mechanisms. Each of these subsets has a smaller number of repeated measures than the original whole data set. For example, in the 12-week smoking cessation study with 174 participants considered in Chapter 4, we have 3 repeated measures per week. By random selecting one of three measures within each week, we obtain a dimension-reduced data subset with a maximum of 12 repeated measures. A test of the intermittent missingness mechanism for the original data may not be possible since the maximum number of repeated measures is 36, which is large compared with the sample size of 174. However, such a test may be feasible in the 12 measure subsets.
We first test whether the intermittent missing mechanism is MCAR without considering the covariates, by using Little’s (1988) likelihood-ratio test. If the test shows that the intermittent missing data are independent of the other repeated measures, then we test CMAR by adding covariates into the testing and using the similar likelihood-ratio method. Otherwise, the mechanism is at least OMAR, and there is no rationale for further testing on CMAR. When CMAR is rejected, then marginal models with GEE and most time-naïve methods are not recommended for analysis. In section 2.2, Little’s Likelihood-ratio test is described and illustrated using several practical data sets.

Test of Dropout Mechanisms

Because special structural characteristics associate with dropouts, there are model-based approaches to test dropout mechanisms (see Table 1). An efficient way to explore the dropout mechanism is to evaluate first for informative dropouts (ID) and only if the test finds no evidence to support this assumption, then proceed to test PDAR vs CDAR/DCAR. The testing of dropout mechanisms is carried out within the framework of MPI where multiple partially-imputed data sets are created and then tested. Finally, these multiple testing results together provide a basis for rejecting or accepting the hypotheses. Similar to testing intermittent missingness mechanisms, we may also resort to a pseudo-bootstrap approach if the number of maximum repeated measure is large.

Note that testing of ID is always associated with a specific model such as the D-K model or the P-M model, as the fitting procedure is specific to the model. We will further discuss differentiating between PDAR and types of ID (e.g., nonignorable random-Coefficient-Based Dropout and nonignorable outcome-based dropout; see Little 1995). According to our experience, in analyzing substance-abuse data sets, testing of PDAR vs. ID is very sensitive to the model choice, and usually the incomplete observed data provide very little information to distinguish between these dropout mechanisms. On the other hand, testing of PDAR vs. CDAR/DCAR is much more robust and feasible in practice. Therefore, in MPI 2.0 and this report, we focus on this latter test. Section 2.3 gives details and illustrations for testing DCAR. Section 4.1 focuses on ID, PDAR, and DCAR within the fitting process of D-K models.

C4. Sensitivity Analysis and Further Assessment of Dropout Mechanisms

In most substance abuse studies, incomplete longitudinal data analysis is sensitive to modeling assumptions. Random effects models (or more generally, mixed models) are often the choice to describe the measurement process should the data be complete, but for incomplete data sets, they are subject to untestable modeling assumptions especially those related to missing data. As we mentioned earlier, some missingness mechanisms (e.g., MCAR and DCAR) might be able to be assessed given only observed data and missingness matrices, but other more
complex missingness mechanisms especially nonignorable ones (e.g., IM or ID) are untestable. When longitudinal models designed to handle ignorable or nonignorable missing data are considered, several choices have to be made. For example, one has to choose between selection and pattern-mixture models, and among various choices for dropout mechanisms.

The sensitivity of selection models has been indicated by Glynn, Laird and Rubin (1986), Draper (1995), and Copas and Li (1997). When the D-K selection model was first proposed, it generated high expectations from the discussants of the paper (Diggle and Kenward 1994). Recently, however, it has been seen that formal tests of dropout mechanisms, especially of informative dropout, although technically possible, should be approached with caution. The statistical community has expressed a growing need for methods that can investigate the sensitivity of the model-based results with respect to the model assumptions (see Nordheim 1984, Little 1994, Rubin 1994, Kenward and Molenberghs 1999, Molenberghs et al. 2001).

In this report, we define a sensitivity analysis as one in which several statistical models along with several assumptions on missingness mechanisms are considered simultaneously for a given incomplete longitudinal data set. This definition encompasses a wide range of useful approaches, such as D-K and P-M models within the framework of MPI and the missingness assumptions listed in Table 1. A two-stage MPI (see Chapter 5) provides the most general framework to perform sensitivity analysis. Most of the modeling strategies are available now through MPI 2.0 for continuous repeated measures.

D. MPI SOFTWARE AND PRACTICAL DATA

D1. MPI 2.0 ---- A Software Package for Incomplete Longitudinal Data Analysis

As a companion to this technical report, the software package named MPI 2.0 has been developed in our SBIR contract project sponsored by the National Institute on Drug Abuse (N43DA-2-5513). MPI 2.0 can be used to do the following tasks for incomplete longitudinal data analysis:

(1) Preliminary Data Exploration with graphical tools, especially on missingness patterns and distribution of the repeated measures.
(2) Maximum Likelihood Estimation using EM-algorithm.
(3) Multiple Partial Imputation using a Gibbs Sampler algorithm.
(4) Testing of MCAR for intermittent missing data.
(5) Testing of DCAR for dropouts.
(6) Fit mixed models for ignorable missing values and dropouts.
(7) Fit D-K selection models to handle nonignorable dropouts.
(8) Fit P-M models to handle nonignorable dropouts.
(9) Fit Shared-Parameter (S-P) models to handle nonignorable intermittent missing values and dropouts.
(10) Consolidate multiple analyses using Rubin’s rule.

MPI 2.0 demonstrates the potential to be developed into a commercial stand-alone software package for incomplete longitudinal analysis. This technical report, also titled “Guide for Statistics”, is used as a companion to the MPI 2.0 User Manual and the software itself.

D2. Practical Data Sets ---- Smoking Cessation in Methadone Maintenance

For illustration purposes in this report, we applied most of the above incomplete longitudinal data analysis techniques to data obtained from a smoking cessation study directed by Dr. Shoptaw.

The aim of the study is to evaluate relapse prevention (RP) and contingency management (CM) for optimizing smoking cessation outcomes using nicotine replacement therapy for methadone-maintained tobacco smokers. The study is a 2 (RP vs. No-RP) by 2 (CM vs. No-CM) randomized clinical trial (see Table 3) with repeated measures including thrice-weekly samples of breath (analyzed for carbon monoxide) and weekly self-reported numbers of cigarettes smoked.

<table>
<thead>
<tr>
<th>Treatment*</th>
<th>CM (No)</th>
<th>CM (Yes)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>RP (No)</td>
<td>42</td>
<td>43</td>
<td>85</td>
</tr>
<tr>
<td>RP (Yes)</td>
<td>42</td>
<td>47</td>
<td>89</td>
</tr>
<tr>
<td>Total</td>
<td>84</td>
<td>90</td>
<td>174</td>
</tr>
</tbody>
</table>

Table 3. Sample size in Treatment Groups

The two types of repeated responses variables -- carbon monoxide levels (CO) and numbers of self-reported cigarettes (CIG) are analyzed separately. There are a maximum of 36 and 12 repeated measures respectively for carbon monoxide levels and self-reported number of cigarettes for each participant during the 12-week study period. Both repeated measures have missing values due to nonresponse and dropout. Some explanatory variables or covariates of interest are listed in Table 4. For more information on the data analysis and the study, refer to Shoptaw et al. (2002).
<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response Variables (repeated measures)</td>
<td></td>
</tr>
<tr>
<td>CO1-CO36 = Carbon Monoxide Levels (three times per week)</td>
<td>0 ppm - 88 ppm</td>
</tr>
<tr>
<td>CIG1-CIG12 = Self-reported Number of Cigarettes (once per week)</td>
<td>0-257</td>
</tr>
<tr>
<td>Explanatory Variables</td>
<td></td>
</tr>
<tr>
<td>COND = Treatment condition</td>
<td>1 = No Treatment</td>
</tr>
<tr>
<td></td>
<td>2 = Relapse Prevention (RP) Only</td>
</tr>
<tr>
<td></td>
<td>3 = Contingency Management (CM) Only</td>
</tr>
<tr>
<td></td>
<td>4 = RP + CM</td>
</tr>
<tr>
<td>CM = CM yes-no dummy variable</td>
<td>1 = Yes</td>
</tr>
<tr>
<td></td>
<td>0 = No</td>
</tr>
<tr>
<td>RP = RP yes-no dummy variable</td>
<td>1 = Yes</td>
</tr>
<tr>
<td></td>
<td>0 = No</td>
</tr>
<tr>
<td>SEX = Gender</td>
<td>1 = Male</td>
</tr>
<tr>
<td></td>
<td>2 = Female</td>
</tr>
<tr>
<td>ETHNIC = Ethnicity</td>
<td>1 = Caucasian</td>
</tr>
<tr>
<td></td>
<td>2 = Hispanic</td>
</tr>
<tr>
<td></td>
<td>3 = African-American</td>
</tr>
<tr>
<td></td>
<td>4 = Asian</td>
</tr>
<tr>
<td>AGE = Subject’s age in years</td>
<td>24-65</td>
</tr>
<tr>
<td>PATCHES = # of nicotine patches used</td>
<td>5-90</td>
</tr>
<tr>
<td>BASELINE = CO level at baseline</td>
<td>1-53</td>
</tr>
</tbody>
</table>

Table 4. A subset of variables in the smoking cessation study
CHAPTER 1

MISSINGNESS PATTERNS

In this chapter, we focus on missingness patterns in incomplete longitudinal data sets and discuss their relationship to selection of longitudinal modeling strategies in substance abuse research. In the next chapter, we will focus on missingness mechanisms.

In describing missingness patterns in this chapter, we first divide missing data into two types, intermittent missing data and dropouts, according to their visual pattern of missingness. Then for each type of missing data, we discuss its role in model selection and fitting for main data analysis. The missingness patterns for the smoking cessation data sets are used for illustration purpose.

1.1 Intermittent Missingness vs. Dropout

Figure 1.1 depicts a typical form of a data matrix for repeated measures, where Variables represents repeated responses, Cases indicates subjects or participants, and question marks refer to missing data. In this typical incomplete longitudinal data set for repeated measures, we see that patterns of missingness can be clearly classified into two types: intermittent missingness and dropout (or immature withdraw). Accordingly, the missing values can be divided into two types: intermittent missing data and dropouts (i.e., missing data due to dropout).

Figure 1.1. Intermittent missingness and dropout in a typical incomplete longitudinal data set
Missingness Patterns for the Smoking Cessation Data

In the smoking cessation study for methadone maintained tobacco smokers (Shoptaw et al. 2002), we have two types of repeated measures: repeated CO (carbon monoxide levels; CO1-CO36) and CIG (self-reported number of weekly used cigarettes; CIG1-CIG12). By importing the original excel data sets into MPI 2.0 (see User’s Guide; BayesSoft 2005) we can obtain the following indicator matrices for missingness patterns in Table 1.1 (for CO1-CO36) and Table 1.2 (for CIG1-CIG12).

Table 1.1 shows the missingness patterns for repeated carbon monoxide levels across the original four treatment conditions: Control, RP, CM, and RP+CM. From this table, we see clearly that there are pretty large amount of missing data due to nonresponse (i.e., intermittent missingness) and due to premature termination (i.e., dropout). Within each treatment group, the overall count of intermittent missing values is smaller than that of the dropouts (i.e., missing values due to dropout), but the number of missingness patterns for intermittent missingness is larger than that for dropout.

Table 1.2 depicts the missingness patterns for repeated self-reported number of weekly used cigarettes after the original four treatment conditions are collapsed into two groups, CMs (CM or RP+CM) and No-CMs (Control or RP), depending on whether participants received CM intervention or not. From Table 1.2, we see only few intermittent missing data points. On the other hand, many dropouts are observed in both groups. Such missingness patterns where dropouts dominate intermittent missing data are not uncommon for repeated self-reported data since the collection procedure for these self-reported data is not highly resource demanding. Most participants are able to provide them by answering phone calls or filling certain questionnaire forms, but once they decide to quit the study, missing data due to dropout become unavoidable. When analyzing these self-reported data, we should pay more attention to the distribution and mechanism of dropouts. The collapsing of treatment conditions here is based on the following considerations. From our preliminary analysis using TES (Ling et al. 1997), an aggregation method, interventions involving contingency management (CM or RP+CM) are effective for these methadone-maintained tobacco smokers. Such a preliminary finding is also of the investigators’ special interest.

Generally, the patterns and rates of missingness are determined by the nature of the design and the study population. For biomarker repeated measures, such as urine metabolites data and breath data, which are usually more study-related or intervention-related, both intermittent missing data and dropouts are frequently observed with patterns similar to those observed in Table 1.1, since the data collection is more “expensive” and complicated. Nonetheless, for longitudinal “survey” data such as self-reported numbers, symptoms, or disorders, the rates of missing data are lower and the patterns of missingness are similar to those observed in Table 1.2, so long as these data are not directly related to the
treatments or interventions and can be considered as secondary data in testing the treatment efficacy.

<table>
<thead>
<tr>
<th>Control</th>
<th>RP</th>
<th>CM</th>
<th>RP+CM</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>4</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Note:
(1) 0=intermittent missing, 1=observed, 2=dropout
(2) The first column in each cell of the table (Control, RP, CM, RP+CM) indicates the number of cases or subjects sharing the missingness vector in that row.

Table 1.1. Missingness Patterns for repeated carbon monoxide levels across treatment conditions in the smoking cessation study
Table 1.2. Missingness patterns for repeated self-reported number of cigarettes across collapsed treatment conditions in the smoking cessation study

1.2 Missingness Patterns and Longitudinal Data Analysis

Earlier in section B2 of the Introduction, we mentioned that the distribution of missingness patterns across treatment conditions is crucial in determining appropriate modeling strategies for data analysis. Here, we use the repeated CO levels in the smoking cessation study as an example to illustrate this point more clearly.

In applying time-naïve methods where repeated measures are summarized into one-dimension composite scores or indices, there is usually an underlying or default assumption associated with the missingness mechanism. For example, in applying TES to the breath data in the smoking cessation study, we first dichotomize the CO levels into either “clean” (if CO > 8 ppm) or “dirty” (if CO ≤ 8 ppm) samples. Then, we count the number of “clean” samples during the 12-week study for each participant. Finally, t-tests or ANOVA can be used to compare the average number of clean samples across treatment conditions. This TES method implicitly assumes that missing data (i.e., breath samples) are “dirty”. This missing-as-dirty assumption is often considered as conservative in testing treatment efficacy since in many studies missingness rates in treatment groups are higher than those in control groups. Nonetheless, in the smoking cessation study, we should be careful in applying TES or explaining the TES-based analysis results.

Table 1.3 lists the overall missingness rates for the four treatment groups in the smoking cessation study for the breath samples (i.e., the percentage of both intermittent missing data and dropouts within each data matrix corresponding to each treatment condition). The CM group has the lowest percentage of missing data, 17.8%. If our interest is to compare CM and Control using TES, for
example, then the missing-as-dirty assumption is no longer conservative. It would be expected that the TES-based analysis would favor CM instead of Control by imputing more dirty samples in the Control group.

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>RP</th>
<th>CM</th>
<th>RP+CM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Missingness Rate</strong></td>
<td>20.9%</td>
<td>20.5%</td>
<td>17.8%</td>
<td>21.7%</td>
</tr>
</tbody>
</table>

**Table 1.3.** Overall missingness rates for carbon monoxide levels within each treatment group in the smoking cessation study

Overall missingness rate only reflects one aspect of a missingness pattern matrix. It is an important factor that should be considered in applying time-naïve data analysis, but there are other patterns of missingness that should also be considered. Suppose in an imaginary study, we are still interested in comparing CM and Control conditions, and in the first 6-week study period there are more missing data in the CM group (100%) than in the Control group (0%), while in the second 6-week study period there are less missing data in the CM group (0%) than in the Control group (100%). By comparing the overall missingness rates between the two groups, we find no difference; both are equal to 50%. However, the t-test using TES actually compares first 6 weeks of Control group data with the second 6 weeks CM group data, which would not be recommended. In this extreme case, retention analysis or survival analysis would offer a complementary tool to TES.

When applying advanced longitudinal models (e.g., marginal models, random effects models, and transition models), the distribution of missingness patterns across treatment conditions becomes less important. So long as the missingness mechanisms are satisfied, the appropriate longitudinal models would be less sensitive to missingness patterns than time-naïve models. However, in many studies, it is difficult to make correct assumptions about the missingness mechanisms, and data-based mechanism assessments (see Chapter 2) are heavily affected by the missingness patterns. Therefore, it is still essential to perform a diagnosis of the distribution of missingness patterns across treatment conditions to ensure that they are not too unbalanced.

### 1.3 Overall Missingness Patterns Across Treatment Conditions

We discuss two approaches to examining missingness patterns: a formal testing approach, and an informal graphical comparison.

#### 1.3.1 Graphical Assessment of Missingness Patterns

Graphical missingness pattern assessment uses plots or tables similar to Figure 1.2-3 and Table 1.1-2. Tables 1.1 and 1.2, produced by MPI 2.0, depict the missingness rates (%) for the 36 repeated CO levels (CO1-CO36) and 12 self-reported numbers of weekly cigarette usage across treatment conditions. Figure
1.2 and 1.3 shows the longitudinal missingness rates (%) for the two types of measures on each time point. From Table 1.1 and Figure 1.2, we see little difference in the missingness patterns across treatment conditions. However, from Table 1.2 and Figure 1.3, we notice some differences in the missingness patterns. There are more subjects with earlier dropout times in the CMs group than in the No-CMs group.

**Figure 1.2.** Percentage of missing data (%) on repeated carbon monoxide levels across treatment conditions (Group: 1=Control, 2=RP, 3=CM, 4=RP+CM)

![Graph showing missing data percentage for carbon monoxide levels](image1)

**Figure 1.3.** Percentage of missing data (%) on self-reported number of weekly used cigarettes across treatment conditions (Group: 1=CMs, 2=No-CMs)

![Graph showing missing data percentage for cigarettes](image2)

### 1.3.2 Statistical Testing of Missingness Patterns

For formal testing of the distribution of missingness across treatment conditions, we view the missingness status (0=intermittent missing, 1=observed, 2=dropout) as a special type of 3-category repeated measures data. Then any type of longitudinal model can be applied to fit these categorical repeated measures and test whether treatment condition is a significant predictor. Since the indicator for missingness status is a categorical variable, when apply mixed models, we should use generalized linear mixed models (GLMMs; Searle and McCulloch 2001). Other models, such as multinomial transition models or marginal models using
GEE, can also be used. To develop and implement each of these modeling strategies into our own software package would require a substantial amount of effort. Considering that these modeling options have been implemented into many popular software packages, MPI 2.0 does not include this functionality.

The following SAS code fits a cumulative logit-regression model in testing the hypothesis of evenly distributed missingness patterns across treatment conditions (COND) for the breath data in the smoking cessation study, when the missingness status is defined as above with 3 categories. The corresponding model fitting results are listed in Table 1.4.

```sas
proc genmod data=GEEdata1;
   class ID T COND;
   model COLEVEL= COND Time COND*Time / dist=multinomial type3;
   contrast 'RP EFFECT' COND -1 1 -1 1;
   contrast 'CM EFFECT' COND -1 -1 1 1;
   repeated subject=ID / within=T;
run;
```

From Table 1.4, we see that formal testing using the 3-category GEE marginal model provides little evidence to reject the hypothesis that the missingness patterns for breath samples in the smoking cessation study are similar across the 4 treatment conditions. **Time**, which is coded 1-36, is a significant factor (p-value<.0001) in predicting the change of missingness rate over the 12-week study period. There is no significant **COND** effect (p-value=0.136) or **Time*COND** effect (p-value=.9175). The p-values for the post-hoc tests of CM and RP effects are also both higher than 0.05.

We now perform an additional analysis where we classify the missingness status as binary: 1=observed, 0=missing (either intermittent or dropout). The following SAS code is used to fit a GEE marginal model (link function=logistic). The results are shown in Table 1.5, from which we still observe little evidence to reject the hypothesis that missingness patterns are evenly distributed across treatment conditions when intermittent missing data are not differentiated from dropouts.

```sas
proc genmod data=GEEdata2;
   class ID T COND;
   model COLEVEL= COND Time COND*TIME / dist=binomial type3;
   contrast 'RP EFFECT' COND -1 1 -1 1;
   contrast 'CM EFFECT' COND -1 -1 1 1;
   repeated subject=ID / type=AR(1) within=T;
run;
```
### Analysis Of GEE Parameter Estimates (Empirical Standard Error Estimates)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Error</th>
<th>Limits</th>
<th>Z Pr &gt;</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept1</td>
<td>-2.3298</td>
<td>0.1948</td>
<td>-2.7116</td>
<td>-1.9481</td>
<td>-11.96</td>
</tr>
<tr>
<td>Intercept2</td>
<td>2.6356</td>
<td>0.2122</td>
<td>2.2197</td>
<td>3.0514</td>
<td>12.42</td>
</tr>
<tr>
<td>COND 1</td>
<td>0.4123</td>
<td>0.3254</td>
<td>-0.2253</td>
<td>1.0504</td>
<td>1.27</td>
</tr>
<tr>
<td>COND 2</td>
<td>0.6274</td>
<td>0.2706</td>
<td>0.0971</td>
<td>1.1577</td>
<td>2.32</td>
</tr>
<tr>
<td>COND 3</td>
<td>0.4213</td>
<td>0.2698</td>
<td>-0.1076</td>
<td>0.9501</td>
<td>1.56</td>
</tr>
<tr>
<td>COND 4</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>time</td>
<td>-0.0558</td>
<td>0.0103</td>
<td>-0.0759</td>
<td>-0.0357</td>
<td>-5.43</td>
</tr>
<tr>
<td>time*COND1</td>
<td>-0.0021</td>
<td>0.0172</td>
<td>-0.0359</td>
<td>0.0288</td>
<td>-0.12</td>
</tr>
<tr>
<td>time*COND2</td>
<td>-0.0062</td>
<td>0.0179</td>
<td>-0.0412</td>
<td>0.0288</td>
<td>-0.35</td>
</tr>
<tr>
<td>time*COND3</td>
<td>-0.0111</td>
<td>0.0168</td>
<td>-0.0440</td>
<td>0.0218</td>
<td>-0.66</td>
</tr>
<tr>
<td>time*COND4</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
</tbody>
</table>

Score Statistics For Type 3 GEE Analysis

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>COND</td>
<td>3</td>
<td>5.54</td>
<td>0.1360</td>
<td></td>
</tr>
<tr>
<td>time</td>
<td>1</td>
<td>41.97</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>time*COND</td>
<td>3</td>
<td>0.51</td>
<td>0.9175</td>
<td></td>
</tr>
</tbody>
</table>

Contrast Results for GEE Analysis

<table>
<thead>
<tr>
<th>Contrast</th>
<th>DF</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>RP EFFECT</td>
<td>1</td>
<td>0.24</td>
<td>0.6220</td>
<td>Score</td>
</tr>
<tr>
<td>CM EFFECT</td>
<td>1</td>
<td>2.16</td>
<td>0.1417</td>
<td>Score</td>
</tr>
</tbody>
</table>

### Table 1.4. Fitting of a GEE marginal model with 3-category missingness status

(0=intermittent missing, 1=observed, 2=dropout)

### Analysis Of GEE Parameter Estimates (Empirical Standard Error Estimates)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Error</th>
<th>Limits</th>
<th>Z Pr &gt;</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.2213</td>
<td>0.2307</td>
<td>1.7691</td>
<td>2.6735</td>
<td>9.63</td>
</tr>
<tr>
<td>COND 1</td>
<td>0.0223</td>
<td>0.3625</td>
<td>-0.6881</td>
<td>0.7327</td>
<td>0.06</td>
</tr>
<tr>
<td>COND 2</td>
<td>0.2494</td>
<td>0.4244</td>
<td>0.0496</td>
<td>1.7177</td>
<td>2.08</td>
</tr>
<tr>
<td>COND 3</td>
<td>0.8837</td>
<td>0.4256</td>
<td>0.0496</td>
<td>1.7177</td>
<td>2.08</td>
</tr>
<tr>
<td>COND 4</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>time</td>
<td>-0.0455</td>
<td>0.0677</td>
<td>-0.0507</td>
<td>0.0304</td>
<td>-5.89</td>
</tr>
<tr>
<td>time*COND1</td>
<td>0.0002</td>
<td>0.0130</td>
<td>-0.0254</td>
<td>0.0237</td>
<td>0.01</td>
</tr>
<tr>
<td>time*COND2</td>
<td>0.0153</td>
<td>0.0358</td>
<td>0.0163</td>
<td>-0.073</td>
<td>0.73</td>
</tr>
<tr>
<td>time*COND3</td>
<td>0.0288</td>
<td>0.0515</td>
<td>0.0585</td>
<td>0.0008</td>
<td>-1.91</td>
</tr>
<tr>
<td>time*COND4</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
</tbody>
</table>

Score Statistics For Type 3 GEE Analysis

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>COND</td>
<td>3</td>
<td>5.74</td>
<td>0.1247</td>
<td></td>
</tr>
<tr>
<td>time</td>
<td>1</td>
<td>42.38</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>time*COND</td>
<td>3</td>
<td>3.98</td>
<td>0.2634</td>
<td></td>
</tr>
</tbody>
</table>

Contrast Results for GEE Analysis

<table>
<thead>
<tr>
<th>Contrast</th>
<th>DF</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>RP EFFECT</td>
<td>1</td>
<td>1.38</td>
<td>0.2402</td>
<td>Score</td>
</tr>
<tr>
<td>CM EFFECT</td>
<td>1</td>
<td>1.17</td>
<td>0.2793</td>
<td>Score</td>
</tr>
</tbody>
</table>

### Table 1.5. Fitting of a GEE marginal model with binary missingness status

(0=intermittent missing or dropout, 1=observed)
For the self-reported cigarettes numbers, since there are only a few intermittent missing data points, we are unable to fit a 3-category logit model as we did for the CO levels. Using binary classification of missingness status, we find no significant difference in the overall missingness patterns across the two collapsed treatment conditions No-CMs vs. CMs (see Table 1.6).

### Analysis Of GEE Parameter Estimates (Empirical Standard Error Estimates)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Error</th>
<th>95% Confidence Limits</th>
<th>Z</th>
<th>Pr &gt;</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.0798</td>
<td>0.2851</td>
<td>2.5211 - 3.6385</td>
<td>10.80</td>
<td>&lt; .0001</td>
<td></td>
</tr>
<tr>
<td>CM 0</td>
<td>0.1144</td>
<td>0.5281</td>
<td>- 0.9206 - 1.1493</td>
<td>0.22</td>
<td>0.8285</td>
<td></td>
</tr>
<tr>
<td>CM 1</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000 - 0.0000</td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>time</td>
<td>-0.1689</td>
<td>0.0223</td>
<td>-0.2127 - 0.1252</td>
<td>-7.57</td>
<td>&lt; .0001</td>
<td></td>
</tr>
<tr>
<td>time*CM 0</td>
<td>-0.0059</td>
<td>0.0467</td>
<td>-0.0975 - 0.0857</td>
<td>-0.13</td>
<td>0.8997</td>
<td></td>
</tr>
<tr>
<td>time*CM 1</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000 - 0.0000</td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
</tbody>
</table>

### Score Statistics For Type 3 GEE Analysis

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>CM</td>
<td>1</td>
<td>0.07</td>
<td>0.7979</td>
</tr>
<tr>
<td>time</td>
<td>1</td>
<td>33.31</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>time*CM</td>
<td>1</td>
<td>0.02</td>
<td>0.8819</td>
</tr>
</tbody>
</table>

**Table 1.6** Fitting of a GEE marginal model with binary missingness status (0=intermittent missing or dropout, 1=observed)

### 1.4 Intermittent Missingness Patterns Across Treatment Conditions

We use the same set of statistical tests to examine the binary indicator of intermittent missingness, coded “1” (observed) or “0” (intermittent missing), with dropouts treated as missing values. Then any appropriate longitudinal models for these “partially observed” binary data can be applied to test the hypothesis that there is no significant difference across treatment conditions of the intermittent missingness patterns.

For the breath data in the smoking cessation study, we applied a GEE marginal model with binomial distribution (similar to the one above Table 1.4) to the intermittent binary missingness status. No significant differences were found (see Table 1.7). For the self-reported number of weekly used cigarettes, we do not need to perform this test since the number of intermittent missing values is small.

In many cases, a separate investigation of the intermittent missingness patterns from that on overall missingness patterns might not be necessary. Assessment of the overall missingness patterns is usually enough for purposes of data exploration and analysis. Nevertheless, the distribution of intermittent missingness patterns across treatment conditions should be considered in performing certain analysis such as retention analysis as seen in Section 1.5.
Table 1.7 Fitting of a GEE marginal model with binary intermittent missingness status (0=intermittent missing, 1=observed; dropouts are ignored by the model)

1.5 Dropout Patterns across Treatment Conditions

1.5.1 Retention Analysis and Survival Analysis

In Sections 1.4 and 1.5, the missingness patterns for dropouts and those for intermittent are viewed and coded in the same way, i.e., a vector of 1/0 missingness indicators for each subject in the study. However, as we mentioned in introductory section B1, dropout patterns for each subject can be coded as a scalar indicator, indicating the time point for the subject’s dropout or, equivalently, the length of his stay in the study. In publications on substance abuse research, data analysis involving these scalar dropout patterns is sometimes called retention analysis or attrition analysis. Biostatisticians would use the term survival analysis. Data that measure lifetime or the length of time until the occurrence of an event (e.g., “death”) are called lifetime or survival data. For longitudinal data, here we can view “dropout” as such an event, thus the data can be viewed as survival data.

To many investigators, retention analysis itself provides important inference about treatment efficacy. A medicine or behavioral therapy may be considered effective
if it keeps subjects longer in the study than a placebo or control condition. Some investigators report their research findings based mainly on this type of analysis, although these derived variables are actually secondary to repeatedly collected metabolite biomarker data. To our understanding, in longitudinal substance abuse studies, the repeated biomarker data should be the main target while retention analysis offers an alternative tool for handling problems caused by dropouts.

In performing retention analysis, intermittent missing data can be treated in two ways. For example, if a subject has the following missingness pattern (“*”=observed, “?”=missing), then survival length can be defined as either 11 (the number of all time points before his dropout) or 6 (the number of actual “clinical visits” before dropout). We have seen that it is meaningful to assess intermittent missingness patterns across treatment conditions (see Section 1.4). For each definition of survival length, intermittent missingness patterns should be comparable across treatment conditions. Otherwise, inference will be biased unless strong missing-data assumptions are satisfied.

There are many software packages available for survival analysis; the following survival analysis results are based on the SAS procedure LIFETEST, which computes nonparametric estimates of a survival distribution function called Kaplan-Meier product-limit (Cox and Oakes 1984, Kalbfleisch and Prentice 1980). The log-rank test and the Wilcoxon test (Kalbfleisch and Prentice 1980) can be used to test homogeneity of survival across treatment conditions.

1.5.2 Survival Analysis for Carbon Monoxide Levels in Breath Samples

Figure 1.4 depicts the product estimate of the survival distribution function for carbon monoxide level data in the smoking cessation study. From this plot, we see that, within each treatment group, survival probability continues to drop from 100% at the start point of the study to about 65% at the termination point. Between the 4th week (Dropout Time Point=12) and the 8th week (Dropout Time Point=24), subjects receiving RP+CM have more dropouts (i.e., smaller survival probability) than subjects receiving the other treatments. However, both the log-rank test ($\chi^2 = 0.117$, p-value=0.99) and the Wilcoxon test ($\chi^2 = 0.065$, p-value=0.995) provide no conclusive evidence to reject the homogeneity hypothesis.
1.5.3 Survival Analysis for Self-Reported Number of Cigarettes

For the self-reported number of cigarettes in the smoking cessation data, the product-limit estimated survival function is depicted in Figure 1.5, from which we see that subjects receiving interventions involving contingency management (CM or RP+CM = Cams) have a higher dropout rate (i.e., lower survival probability) throughout the 12-week study period. This is consistent with visual observation of the dropout patterns in Table 1.2. Once again, the log-rank test ($\chi^2 = 0.049$, p-value=0.82) and ($\chi^2 = 0.127$, p-value=0.72) suggest that the difference is not significant between the treatment conditions with CM (i.e., CMs) and those without CM (i.e. No-CMS).
SUMMARY FOR CHAPTER 1

In this chapter, we used data sets from a smoking cessation study to illustrate how to perform preliminary analysis of the missingness patterns and how to investigate their distributions across treatment conditions. Since unbalanced distributions of intermittent missingness patterns and/or dropouts can potentially affect our longitudinal data analyses especially those adopting time-naïve models, we should have a systematic way of testing or assessing various forms of homogeneity hypotheses. Graphical diagnoses usually provide consistent results with formal statistical tests, but the latter methods are more flexible and provide p-values. From the above preliminary analyses, we found little evidence to support heterogeneity across treatment conditions in missingness patterns for the smoking cessation study. In selecting longitudinal modeling strategies, we prefer longitudinal models (e.g., mixed models, marginal models, or transition models) to time-naïve methods (e.g., TES, JSP, etc.) since the latter type of models require stronger assumptions on missingness patterns (as discussed in this Chapter) and missingness mechanisms (which will be discussed in the next Chapter).
CHAPTER 2

MISSINGNESS MECHANISMS

Properties of methods for handling missing data are strongly influenced by assumptions made about the missingness mechanism. It is important to consider the nature of this mechanism in a particular application in order to make unbiased parameter estimation and valid inference. For subject $i$ with incomplete repeated measures $Y_i$, a formal definition of missingness mechanism is given in the introductory section B3, that is,

$$\text{Missingness Mechanism} = P(R_i | Y_i, \psi),$$

(2.1)

where $R_i$ is the vector of missingness indicators, and $\psi$ is the associated missingness parameters. To better understand these definitions, consider the following examples.

Suppose in a longitudinal study, we repeatedly measure daily systolic blood pressures (SBPs) during one week from Monday to Friday. If there were no missing data, we would observe the following 5 SBPs for Mr. Armstrong:

$$Y_i = (120\text{mmHg}, 130\text{mmHg}, 180\text{mmHG}, 125\text{mmHG}, 140\text{mmHG}).$$

(2.2)

Unfortunately, Mr. Armstrong did not go to the clinic on Wednesday. As a result, we missed the BPS of 180mmHg for Wednesday, and correspondingly the missingness pattern is $R_i=(1,1,0,1,1)$. There may be many reasons for this missing value. For example, he did not go because it snowed heavily that day or he had a schedule confliction. Or perhaps he felt dazed and weak, because his blood pressure was high (180mmHg). This is a common type of missingness mechanism, missingness caused by an extreme value of the quantity under study. The missing value happens because it is higher than a censoring level, $\psi$, which is either unknown or known, e.g. $\psi=150\text{mmHg}$. Ideally, when performing data analysis, such censoring mechanisms should be modeled in addition to the modeling of the repeated SBPs.

Recall that mechanisms for intermittent missingness (see Table 1) can be roughly divided into three types: missing completely at random (MCAR), missing at random (MAR), and informative missingness (IM). Using Mr. Armstrong’s data, Table 2.1 gives examples for these missingness mechanisms. In the case of IM, the Wednesday missingness depends on the underlying missed value from the same day, 180mmHg. In the case of MAR, the Thursday missingness does not depend on the Thursday data (i.e., 125mmHg), but it depends on data from previous observed days, such as Wednesday’s 180mmHg. As an MCAR case, the
Tuesday missingness does not depend on the Tuesday data (130mmHg), nor does it depend on any other day’s data (perhaps the visit was missed due to heavy snowfall).

<table>
<thead>
<tr>
<th>Missingness Mechanisms</th>
<th>Repeated Measures ((Y_i))</th>
<th>Missingness Indicator ((R_i))</th>
</tr>
</thead>
<tbody>
<tr>
<td>IM</td>
<td>(120, 130, ?, 125, 140)</td>
<td>(1, 1, 0, 1, 1)</td>
</tr>
<tr>
<td>MAR</td>
<td>(120, 130, 180, ?, 140)</td>
<td>(1, 1, 1, 0, 1)</td>
</tr>
<tr>
<td>MCAR</td>
<td>(120, ?, 180, 125, 140)</td>
<td>(1, 0, 1, 1, 1)</td>
</tr>
</tbody>
</table>

**Table 2.1 Possible missingness mechanisms for intermittent missingness**

Similarly, in Table 2.2 we give examples for dropout mechanisms (see Table 1), which are generally classified as DCAR, DAR, or ID. The explanations for them are parallel to those for intermittent missingness mechanisms.

<table>
<thead>
<tr>
<th>Missingness Mechanisms</th>
<th>Repeated Measures ((Y_i))</th>
<th>Missingness Indicator ((R_i))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informative Missing</td>
<td>(120, 130, ?, ?, ?)</td>
<td>(1, 1, 0, 0, 0)</td>
</tr>
<tr>
<td>DAR</td>
<td>(120, 130, 180, ?, ?)</td>
<td>(1, 1, 1, 0, 0)</td>
</tr>
<tr>
<td>DCAR</td>
<td>(120, ?, ?, ?, ?)</td>
<td>(1, 0, 0, 0, 0)</td>
</tr>
</tbody>
</table>

**Table 2.2 Possible dropout mechanisms**

Note that all above definitions and examples for missingness mechanisms are given on the individual level (e.g., Mr. Armstrong). We use similar taxonomy for missingness mechanism on the group level. For example, within each treatment group, data are MAR when the distribution of the missingness patterns (i.e., \(R_i\)) does not depend on the missed values (i.e., \(Y^{(o)}\)), but only on the observed ones (i.e. \(Y^{(m)}\)). This would include the case that Mr. Armstrong’s Thursday missingness may not depend on his own data, but does depend on the observed data of somebody else in his treatment group (e.g. the person who gives him a ride to the clinic). Although this might be unusual in practice, it could be seen in community-based studies where participants live close to each other and can influence others attitudes toward the treatment they receive.

In this chapter, we will first explain why certain mechanisms are called ignorable but others are called nonignorable in performing longitudinal data analysis. Then, we describe how to investigate possible missingness mechanisms given only the observed data in a study. Throughout, the smoking cessation data sets are used for demonstration.
2.1 Ignorable and Nonignorable Missingness Mechanisms

As mentioned in section B4 of the Introduction, an appropriate longitudinal modeling strategy should consider both the observed data and missingness patterns by modeling their joint distribution, i.e., \( p(Y_i, R_i | X_i, \theta, \psi) \) for the \( i^{th} \) subject. Correspondingly, the full form of the likelihood function

\[
L_{\psi}(\theta, \psi | Y^{(o)}, R) \propto \prod_{i=1}^{N} \int_{Y^{(m)}} p(Y^{(o)}_i, R_j, Y^{(m)}_i | X_i, \theta, \psi) dY^{(m)}_i \tag{2.3}
\]

should be used for parameter estimation. In most popular software packages such as the MIXED procedure in SAS, however, the longitudinal modeling parameters are estimated using a reduced form of the likelihood function,

\[
L_{\psi}(\theta | Y^{(o)}) \propto \prod_{i=1}^{N} \int_{Y^{(m)}} p(Y^{(o)}_i, Y^{(m)}_i | X_i, \theta) dY^{(m)}_i \tag{2.4}
\]

which we call the observed-data likelihood function. This certainly simplifies the model-fitting procedure and saves computational cost, but may result in estimation bias or inconsistency.

As Rubin (1976) found, when the missingness mechanism for a data set satisfies certain properties, which he called ignorability, meaning

1. the missing data are MAR, and
2. the parameters for the data, \( \theta \), and the parameter for the missingness mechanism, \( \psi \), are “distinct”

then we can safely ignore the missingness mechanism by estimating \( \theta \) from the observed-data likelihood. “Distinct” can be interpreted as “the joint parameter space for \( (\theta, \psi) \) is the Cartesian cross-product of the individual parameter space for \( \theta \) and \( \psi \) from a frequentist’s view,” or “the joint prior distribution \( \pi(\theta, \psi) \) equals to \( \pi_{\theta}(\theta)\pi_{\psi}(\psi) \) from a Bayesian perspective” (Schafer 1997).

When ignorability holds for a missingness mechanism, we say that it is ignorable. It is safe to ignore the missingness mechanism because of the following argument:

\[
P(R, Y^{(o)} | \theta, \psi) \propto \int P(R, Y | \theta, \psi)dY^{(m)}
\]

\[
= \int P(R | Y, \psi)P(Y | \theta)dY^{(m)}
\]

\[
= P(R | Y^{(o)}, \psi)\int P(Y | \theta)dY^{(m)}
\]

\[
= P(R | Y^{(o)}, \psi)P(Y^{(o)} | \theta)
\]

where we see that the full form of likelihood (i.e., \( P(R, Y^{(o)} | \theta, \psi) \), can be partitioned into the product of the likelihood for the missingness pattern (i.e., \( P(R | Y^{(o)}, \psi) \)) and the observed-data likelihood (i.e., \( P(Y^{(o)} | \theta) \)).

From a Bayesian perspective, under ignorability, the observed-data posterior distribution can be defined as

\[
P(\theta | Y^{(o)}) \propto L_{\psi}(\theta | Y^{(o)})\pi_{\theta}(\theta), \tag{2.6}
\]
where \( \pi_{\theta}(\theta) \) is the marginal prior distribution for \( \theta \). This observed-data posterior distribution is derived from

\[
P(\theta \mid R, Y^{(a)}) = \int P(\theta, \psi \mid R, Y^{(a)}) d\psi
\]

\[
\propto P(Y^{(a)} \mid \theta) \pi_{\theta}(\theta) \int P(R \mid Y^{(a)}, \psi) \pi_{\psi}(\psi) d\psi .
\]

(2.7)

\[
\propto L_{\psi}(\theta \mid Y^{(a)}) \pi_{\theta}(\theta)
\]

It is plausible that \( \theta \) and \( \psi \) are distinct to each other in many data sets, for example, the censoring mechanism for Mr. Armstrong’s blood pressure data, where censoring level \( \psi = 180\text{mmHg} \) is distinct from \( \theta = (\mu, \Sigma) \), the mean vector and the covariance matrix of the 5 repeated SBPs. This is more obvious in the MCAR cases, for example, Armstrong’s Tuesday missingness due to snowing, where \( \psi \) can be viewed as the probability of snowing on that day and is clearly unrelated to \( \theta = (\mu, \Sigma) \). In certain studies, such a distinctness requirement may not be so obviously satisfied, but we may still be able to redefine \( \psi \) to make it less dependent of \( \theta \). For example, if the censoring level (e.g., \( \psi = 180\text{mmHg} \)) for Armstrong’s Wednesday blood pressure is positively related to the expectation of his Wednesday blood pressure (e.g., \( \mu = 160\text{mmHg} \)), then we can redefine \( \psi \) as the difference between the censoring level and the expectation level (e.g., \( \psi = 20\text{mmHg} \)).

Therefore, in this technical report, we treat MCAR and MAR as ignorable missingness mechanisms, assuming that the distinctness requirement is satisfied. Among the missingness mechanisms listed in Table 1 (see section B3) for incomplete longitudinal data, we see that MCAR/DCAR, CMAR/CDAR, OMAR/PDAR are ignorable and IM/ID is nonignorable.

### 2.2 Missingness Mechanisms and Selection of Modeling Strategies

For a practical data set, a good understanding of its possible missingness mechanisms is crucial for model selection and model fitting. Before choosing a longitudinal modeling strategy, an investigation of the missingness mechanism is an essential part of the preliminary data exploration or analysis.

Let us take the following hypothetical data in Table 2.3 as an example, where six cocaine-dependent gay men from West Hollywood participated in a baclofen-placebo trial and provided at most five urine samples with Benzoylecognine level being measured. There are missing measures within each group.

<table>
<thead>
<tr>
<th>Placebo</th>
<th>Baclofen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject 1</td>
<td>500 400 300 200 100</td>
</tr>
<tr>
<td>Subject 2</td>
<td>500 400 300 200 100</td>
</tr>
<tr>
<td>Subject 3</td>
<td>500 400 300 ---- ----</td>
</tr>
<tr>
<td>Subject 4</td>
<td>500 400 300 200 100</td>
</tr>
<tr>
<td>Subject 5</td>
<td>500 400 300 200 100</td>
</tr>
<tr>
<td>Subject 6</td>
<td>800 700 600 ---- ----</td>
</tr>
</tbody>
</table>

Table 2.3. Benzoylecognine levels from a hypothetical baclofen-placebo study
For data analysis in evaluating the treatment effects, we compare the change of the mean Benzoylecognine levels across treatment groups. Based on all observed data, the mean vector for placebo is calculated as $\mu_0 = (500, 400, 300, 200, 100)$ and the mean vector for baclofen is $\mu_1 = (600, 500, 400, 200, 100)$. Since the drop from 600 to 100 is larger than the drop from 500 to 100, this simple available data analysis ignoring the missing data favors baclofen over placebo. Nonetheless, this analysis is valid only if the missing data are MCAR, i.e. if the expectations of the missing data in the two groups have nothing to do with other observed data. In this case, the missing values are assumed to have expectation equal to the mean for that time point.

Taking a further look at the data, we see that subject 6 has higher observed Benzoylecognine levels than subject 3. This suggests that the missing values for subject 6 should be higher than the missing values for subject 3, if the missingness mechanism is MAR instead of MCAR. Realistically, if we are sure that the missing data are MAR and all the subjects have the same linear trend of their repeated measures, the missing data on Subjects 3 and 6 might be respectively predicted as (200, 100) and (500, 400) based on their observed data and the data of other subjects. Using this imputation with the MAR assumption, the comparison will show no difference between the two treatment conditions. In other words, available data analysis under the MCAR assumption is actually associated with a bias favoring baclofen.

Another possible case for the missing data is IM (or nonignorable missing). In most studies, it is usually hard to determine the real missingness mechanisms. But let suppose that investigators for this baclofen-placebo study trust that missing data happens because they are censored at the level of 1000. Under this conjecture, the missing data are all higher than 1000. Imputing the missing data values as equal to exactly 1000, we have $\mu_0 = (500, 400, 300, 466, 400)$ and $\mu_1 = (600, 500, 400, 466, 400)$. If treatment effect is defined as the drop rate of Benzoylecognine levels from the first time point to the last time point, that is, (500-400)/500=0.2 for placebo and (600-400)/600=0.3 for baclofen. Thus, baclofen has a higher effect size (0.3-0.2=0.1) than placebo. Using the same definition for effect size, available data analysis favors baclofen over placebo with an effect-size difference of (600-100)/600-(500-100)/500 =0.03, smaller than 0.3. In this IM case, the estimate from available data analysis is still biased but this time it is “conservative” in the sense that the bias is on the direction that favors placebo. Since the censored missing values should be higher than 1000, we can impute them by any levels higher than 1000. It is easy to verify that as long as the missing values are close to each other, available data analysis is consistent with the IM-based imputation analysis in supporting baclofen over placebo.

From the above analysis, we clearly see that analysis plans for incomplete longitudinal data are highly dependent on the assumptions on missingness.
mechanisms. Although in the above baclofen-placebo illustration, we only focused on general missingness mechanisms, similar relationships to data analysis are found for intermittent missingness mechanisms and dropout mechanisms in practical longitudinal data sets.

As we will discuss more fully in Section 3.1, there are generally three ways to handle missing data:

1. **complete case analysis**, where subjects with any missing data are excluded from consideration in model fitting;
2. **analyzing as incomplete**, or **observed-data analysis**, where all the available data are considered in model fitting (this is in contrast to complete case analysis);
3. **imputation method**, where missing data are replaced by some “plausible” values and then standard complete-data analysis is performed.

When performing longitudinal data analysis using either time-naïve models or advanced longitudinal models, all of the above methods can be used. According to our experience and literature reviews of substance abuse research, we make the following observations:

1. complete case analysis is rarely seen in research reports and proposals because **intention to treat analysis** is now accepted by most of the clinical trial community as the standard for data analysis, and it is required for FDA review. Intention to treat analysis requires that all subjects randomized into the study should be included in the analysis.

2. In many traditional time-naïve models, the handling of missing data usually implies an imputation method with “missing-as-dirty” assumption. For example, a composite score called JSP by Ling and his colleagues (1997) is defined as the number of “clean” urine samples among all the observed repeated measures on each person. The missing urine samples are thus implicitly considered to be “dirty” samples. These time-naïve approaches could seriously bias analyses, sometimes with the unintended consequence of favoring an experimental condition over a control condition (Yang and Shoptaw 2005). Earlier in Chapter 1, we have mentioned that these approaches require very strong assumptions on the missingness patterns.

3. When applying marginal models with quasi-likelihood methods such as GEE, the intermittent missing data should be MCAR and dropouts should be DCAR, in order to obtain unbiased parameter estimation. For missingness mechanisms that are not MCAR or DCAR, modified GEE algorithms such as the one proposed by Paik (1997) should be used.

4. Transition models based on Markov Chains require certain types of MAR (e.g., CMAR, OMAR, CDAR, and PDAR) assumptions where missingness on
the present measure depends only on the preceding observed measures (i.e., the history data). In many practical settings, this left-side data dependent MAR is plausible.

(5) Missingness assumptions for random effects models or mixed models are less rigorous, including any type of MAR (e.g., CMAR, OMAR, CDAR, and PDAR). This modeling strategy offers the most flexibility in developing advanced longitudinal models such as random-effects selection models and random-effects pattern-mixture models which are designed potentially to handle informative missing data (e.g., IM and ID).

(6) Advanced models that combine several techniques of the above modeling strategies have appeared in the recent statistical literature; for example, see articles by Diggle and Kenward (1994), Nordheim (1984), Little (1994), Rubin (1994), Kenward and Molenberghs (1999), West and Harrison (1999), Molenberghs et al. (2001). Some of these models can be used to deal with informative missing data, or provide a way to perform sensitivity analysis. In this technical report, we focus on Diggle and Kenward’s selection model (i.e., D-K model) and Little’s pattern-mixture model (i.e., P-M model). West and Harrison’s dynamic linear models for incomplete longitudinal data in substance abuse research will be a topic for future research.

Since assumptions on missingness are critical for valid statistical inference, there is strong incentive to investigate the underlying mechanisms that explain intermittent missing data and dropouts. This process is ideally conducted with new evidence from outside of the observed data. For instance, might review historical records to understand reasons behind participants’ non-responses or conduct follow-up phone calls to investigate subjects’ dropout. Unfortunately, these methods are often resource intensive and may be ethically indefensible. Independent of these, statistical analysis can provide data-based evidence for missingness mechanisms.

2.3 Tests of MCAR for Intermittent Missing Data

2.3.1 Data-Based Evidence Algorithms for Testing MCAR

Based only on information from observed data, it may be impossible to determine whether the mechanism for intermittent missingness is nonignorable, however it is often possible to distinguish between MCAR and other mechanisms, i.e., MAR or NM, based on the observed data. In a multivariate normally distributed data set, a simple way of assessing MCAR is to compare the means of observed values of each variable between groups defined by whether other variables are missing or not. However, when there are many variables all possibly with missing data, a difficulty in applying this method is that many related test statistics are used in
this process. To solve this multiple-comparison problem, Little (1988) proposed the following likelihood ratio statistic for testing MCAR in multivariate continuous data sets:

\[
d^2 = \sum_{p=1}^{P} m_p (\bar{y}_{p}^{(o)} - \hat{\mu}_{p}^{(o)}) (\hat{\Sigma}_{p}^{(o)})^{-1} (\bar{y}_{p}^{(o)} - \hat{\mu}_{p}^{(o)})^T
\]

where \( P \) = the number of unique missingness patterns,
\( m_p \) = number of subjects with \( p^{th} \) pattern (i.e., in the \( p^{th} \) strata),
\( \hat{\mu}_{p}^{(o)} \) = M.L.E. of the mean vector in the \( p^{th} \) strata,
\( \hat{\Sigma}_{p}^{(o)} \) = M.L.E. of the variance-covariance matrix in the \( p^{th} \) strata.

Note that in this definition, the data set are stratified into subgroups or strata according to the missingness patterns. Asymptotically, this statistic has a chi-square distribution with \( df = \sum_{p=1}^{P} K_p - K \) degrees of freedom, that is,

\[
d^2 \sim \chi^2_{df}
\]  

(2.8)

where \( K_p \) is the number of observed variables in the \( p^{th} \) strata, and \( K \) is the number of all variables in this multivariable continuous data set. Park and Davis (1993) extended this test to multivariate categorical data.

Since the data matrix for repeated measures is itself a multivariate normal data set (within the scope of this technical report), Little’s likelihood ratio statistic can be directly applied when testing MCAR for intermittent missing data. Nonetheless, there is another difficulty associated with such a test for longitudinal data where the dropouts need to be considered. When the number of participants with dropouts is small, we may safely exclude them when conducting the test of MCAR. Otherwise, we suggest the following restricted pseudo-bootstrap algorithm,

| **Table 2.4** A pseudo-bootstrap algorithm for testing MCAR |

The p-values from this pseudo-bootstrap can jointly be used to assess the MCAR hypothesis. If intermittent missing data in the original data set are indeed MCAR, most of the p-values should be larger than a pre-specified significant level, say
\( \alpha = 0.05 \). Otherwise, if most of the p-values are smaller than this significance level, we have strong evidence for rejecting the MCAR hypothesis. As we mentioned earlier in section C2 of the Introduction, in substance abuse data sets, \( K \) is usually for direct application of Little’s test because the number of intermittent missingness patterns will be too large. By subselecting a data set with a smaller number of \( J \leq K \) repeated measures, the number of intermittent missingness patterns becomes much smaller.

### 2.3.2 Test of MCAR for Intermittent Missingness in the Smoking Cessation Study

To test whether intermittent missing CO levels were MCAR in the smoking cessation study, we applied the above pseudo-bootstrap algorithm. We divided the 174 participants into 4 groups depending on condition assignment. Then for each group, we randomly selected one of the three repeated CO measurements within each week for each person. This single CO level approximately represents the smoking status during that week. For each selected subset with 12 repeated measures, we dropped all the cases with dropouts and then applied Little’s likelihood test for MCAR. For each of the 4 groups, we resampled 100 subsets and obtained 100 p-values in applying Little’s MCAR test. Figure 2.1 shows the histograms of these p-values for each group. From the four histograms, we clearly observe that the hypothesis of MCAR could absolutely be rejected for intermittent missingness in the groups receiving CM interventions. But for intermittent missing data in the groups without CM, there is less evidence against MCAR. A trend clearly seen from the plots is that from Control, to RP, to CM, and to RP+CM, the assumption of MCAR becomes increasingly less probable.

![Histograms of p-values](image)

**Figure 2.1** Histograms of p-values in testing MCAR for intermittent missing CO levels in the smoking cessation study.
We performed these tests again adding four covariates (Ethnicity, Age, Number of Patches, and Baseline CO) to the 12-measure data sets. We selected these covariates since preliminary analysis showed they are strongly related to the repeated CO levels. No significant differences were observed from the results depicted in Figure 2.1. Therefore, we conclude that intermittent missing data in the treatment groups with contingency management cannot be viewed as MCAR in performing longitudinal analysis.

A plausible explanation for the findings in Table 2.1 might be that those in the CM groups were reluctant to provide breath samples when they smoked prior to their clinical visit or when they suspected that their CO level would exceed the criterion used to determine tobacco use or nonuse, and thus they would not receive their incentive. When receiving contingency management, participants are given high-valued incentives, such as vouchers for food, if their CO level indicates no recent tobacco use. Without additional evidence beyond that in the observed data, it is better to assume that intermittent missing data are at least MAR in CM and RP+CM groups and possibly in the RP group. The assumption of MCAR is too liberal when the goal is to conduct unbiased analyses.

Since there are only a few intermittent missing data points for self-reported number of cigarettes, we did not perform the test of MCAR for them.

### 2.4 Assessing Dropout Mechanisms within the Framework of MPI

#### 2.4.1 MPI-Multiple Partial Imputation

MPI refers to multiple partial imputation, a Monte Carlo technique, which performs multiple imputation (Rubin 1987) only for the intermittent missing data and leaves the dropouts as they are. For details on why and how to perform MPI, please see Chapter 3.

When our goal is testing of dropout mechanisms, MPI offers a useful tool by producing partially imputed data sets where all missing data are due to dropout so that we can more directly investigate dropout mechanisms. In the following discussion, we assume that all intermittent missing data have been imputed and each testing strategy is performed on one of the multiply partially imputed data sets.

#### 2.4.2 Data-Based Evidence for Assessing Dropout Mechanisms

A diagnostic tool for determining reasons for dropout is via exploratory data plots (Verbeke and Molenberghs 2000). Because special structural characteristics associate with missing data caused by dropouts, there are also model-based testing approaches.
Diggle (1989) proposed a method to test whether dropouts within a treatment group are DCAR. The principle idea is that, at each time point within each treatment group, the subjects who drop out are a random sample from all the subjects who do not drop out, if the dropouts are DCAR. More specifically, let us define

\[ h(y, i) = \sum_{k=1}^{j-1} w_k y_{ik} \quad (i = 1, ..., N) \]  

(2.9)
as a linear function of the preceding observations for \( y_{ij} \) on dropout time \( t_j \) (\( j = 2, ..., K \)) with some weights \( w_k \left( \sum_{k=1}^{j-1} w_k = 1 \right) \). Denoting \( R_j \) and \( r_j \) respectively as the number of subjects who do not drop out and the number of subjects who drop out at \( t_j \), then asymptotically, we have

\[ \bar{h}_j = r_j^{-1} \sum_{i=1}^{j-1} h_{ij} \sim N(\overline{H}_j, S_j^2 (R_j - r_j) / (r_j R_j)) \]  

(2.10)

where \( \overline{H}_j = R_j^{-1} \sum_{i=1}^{R_j} h_{ij} \) and \( S_j^2 = (R_j - 1)^{-1} \sum_{i=1}^{R_j} (h_{ij} - \overline{H}_j)^2 \). Correspondingly, a t-test would produce a p-value for testing DCAR at this time point. All the p-values for all dropout time points should be distributed as a uniform \((0,1)\) distribution when overall dropouts are DCAR. Violation of this uniform distribution can be measured by Barnard’s Monte Carlo p-value or the Kolmogorov statistic (Diggle 1989). In applications, we should check that the above testing scheme is not overly sensitive to the choices of \( w_k \). Otherwise, the test results are not particularly trustworthy. We recommend a default specification of \( w_k = 2^{-(j-k)} \).

Similar to the algorithm in Table 2.4, we perform another pseudo-bootstrap algorithm for testing DCAR when the number of repeated measures is relatively large compared to a small sample size. In this algorithm, subsets of the partially imputed data set with smaller numbers of repeated measures are randomly resampled, on which Diggle’s tests are applied.

Use MPI Gibbs Sampler to create multiple, say \( M \), partially imputed data sets \( Y_1, Y_2, ..., Y_M \).

For \( i = 1, ..., M \), do {
    For \( b = 1, ..., B \), do {
        1). Randomly Select a subset \( Y_i^* \) from \( Y_i \) with \( J^* \) repeated measures (\( J^* \leq J \)).
        2). Apply Diggle’s test on \( Y_i^* \) to get a p-value \( p_i^{(b)} \).
    }
}

Table 2.5 A Pseudo-Bootstrap Algorithm for testing DCAR
The set of p-values is then used jointly as evidence to reject or accept the hypothesis that dropouts are DCAR.

When the test results provide strong evidence to reject the assumption of DCAR, we may be able to use other tools for further assessment of the dropout mechanisms. For example, D-K models (Diggle and Kenward 1994) can be used to test whether dropouts are ID or DAR (i.e., CDAR, or PDAR). This will be fully discussed in Chapter 4; see also Yang and Shoptaw (2005).

2.4.3 Testing of DCAR for Breath Samples in the Smoking Cessation Study

In the smoking cessation study, the investigators were more interested in the effect of contingency management. As we did in section 1.1 for self-reported data, we first collapsed the four treatment conditions into two groups: CMs (CM or RP+CM) and No-CMs (Control or RP). Then within each of the two groups, we performed DCAR testing using the algorithm in Table 2.5.

For each data set, four partially imputed versions were created by MPI 2.0 using a Gibbs sampler (see User’s Guide). For each of these partially imputed data sets, we randomly selected 100 subsets, each with 12 repeated measures (once per week). Then, MPI 2.0 performed Diggle’s test on each of these 12-measure subsets to obtain 100 p-values, which are shown in Figure 2.2. When performing Diggle’s test, we set \( \lambda_k \). An example of Diggle’s test for a 12-measure subset including all Monday CO levels is given in the User’s Guide. From Figure 2.2, we observe a trend that is consistent with the MCAR testing results: the assumption of DCAR can be rejected in the CMs groups, but not in the No-CMs groups. There was little variance between imputations, which further supports these beliefs.

The same procedure was applied to the original four groups (Control, RP, CM, and RP+CM). We obtained testing results that are similar to those observed in Figure 2.1 for MCAR testing. Nonetheless, since the sample sizes for the four treatment conditions are smaller, the tests are less powerful than those associated with the two collapsed data sets. Table 2.6 gives an example of the testing p-values based on a single subset for each partially imputed data set. From this table, we see that DCAR becomes less acceptable when the treatment condition changes from Control, to RP, to CM, and to RP+CM. Once again, similar explanations for intermittent missing data apply here as for the dropouts.

Since the testing of DCAR was performed within each treatment group using only repeated measures and there is significant evidence rejecting the assumption of DCAR for participants receiving CM, we believe that the dropout mechanisms are at least PDAR for these participants. In Chapter 4, we perform further analysis to assess whether dropouts are ID or not. Sensitivity analyses will be introduced that potentially can be used to deal with nonignorable dropouts.
Figure 2.2 Histograms of p-values in DCAR testing for dropout CO levels in the smoking cessation study.

Table 2.6 Kolmogorov-test for testing DCAR within treatment conditions in four partially imputed data sets.

2.4.4 Testing DCAR in the Self-reported Data in the Smoking Cessation Study

In a similar manner, we applied Diggle’s test of DCAR for the dropouts in the self-reported number of weekly used cigarettes with treatment condition collapsed into No-CMs and CMs. The histograms of the testing p-values are plotted in Figure 2.3, from which we see very little evidence to reject DCAR for the No-CMs group, but we do observe some evidence (although not strong) for the CMs group.
Figure 2.3 Histograms of p-values from testing DCAR for self-reported data in the smoking cessation study

SUMMARY FOR CHAPTER 2

In this chapter we used several examples to illustrate the taxonomy of missingness mechanisms within the setting of outcome-dependent missingness/dropout and their important role in selecting and fitting proper models for longitudinal data analysis. We proposed two pseudo-bootstrap algorithms based on Little’s likelihood ratio test and Diggle’s test in testing MCAR for intermittent missing data and DCAR for dropouts in the smoking cessation study. They are called “pseudo-bootstrap” instead of “bootstrap” because they resample subsets of repeated measures instead of resampling subjects. In MPI 2.0, a test of MCAR or DCAR is only performed on one data set. To perform these pseudo-bootstrap algorithms in MPI 2.0, we should first resample a certain number of subsets from the original large data set and then perform the MCAR or DCAR test within MPI 2.0 for each of these sub-datasets. In chapters 5 and 6, mechanism of missingness
and dropout will be further discussed under other settings (i.e., pattern-dependent and parameter-dependent missingness/dropout).

As several practical data analyses from the UCLA Integrated Substance Abuse Program (UCLA-ISAP) suggest, for intermittent missingness of repeated biomarker data such as urine samples, breath samples, etc., MCAR is rarely plausible especially in experimental conditions rather than control conditions. For secondary collected repeated measures such as self-reported symptoms on mental health or number of substance uses, intermittent missing data might be plausibly MCAR. These observations are consistent with those seen with intermittent missingness patterns (see Chapter 1). Similar explanations apply here, that is, these secondary data are less study-related and less expensive to collect.

For dropouts, our limited practical experience suggests that the assumption of DCAR or CDAR is seldom true for both biomarker data and secondary survey-like data. This can be explained by the fact that once subjects withdraw from a study we usually lose them forever, and no secondary data can be obtained from them. The reasons behind these withdrawals are usually study-related and cannot be viewed as random factors.

**Covariate-related Missingness Mechanisms**

In this chapter, we discussed missingness mechanisms; we roughly divided these into three categories MCAR (or DCAR), MAR (or DAR), and IM (or ID). We did not specifically consider the classifications that involve covariates such as treatment conditions. These covariate-related mechanisms, i.e., CMAR and CDAR, sometimes provide additional guidance in longitudinal data analysis. For example, marginal models using GEE provide unbiased estimation when the missingness mechanism is up to CMAR and CDAR.

Since we perform MCAR (or DCAR) tests within each treatment group using only observed repeated measures, when these tests show that intermittent missing data (or dropouts) are not MCAR (or DCAR), it is not necessary to test whether they are CMAR (or CDAR). A violation of MCAR (or DCAR) in this situation implies that intermittent missing data (or dropouts) are at least OMAR (or PDAR). Therefore, the original form of marginal models using GEE is not recommended since they are not able to deal with OMAR (or PDAR).

On the other hand, if the within-group test accepts the null hypothesis of MCAR (or DCAR), it is now possible that the intermittent missing data (or dropouts) are CMAR (or CDAR). Usually a further assessment of CMAR (or CDAR) is unnecessary since many longitudinal models, e.g. GEE marginal models, work equivalently well under MCAR (or DCAR) or CMAR (or CDAR). However, other models especially time-naïve models, e.g., TES and JSP, are sensitive to this further distinction.
It is easy to differentiate between MCAR (or DCAR) and CMAR (or CDAR) once we know that, within each treatment condition, missingness patterns are independent of the observed data. To do this, we combine all the within-treatment data on repeated measures into one large data set and add in all covariates of interest. After redefining missingness patterns for this large data set, we can apply Little’s likelihood ratio test or Diggle’s test to assess CMAR vs. MCAR or CDAR vs. DCAR. In applying Little’s likelihood test, we should use the data on covariates as the multivariate normal data \((y)\) with the strata determined by the redefined missingness patterns. Similarly, the \(h(y_{i1},...,y_{i,j})\) in applying Diggle’s testing method are defined for the covariates instead of the repeated measures.

**Dropout Mechanisms and Sensitivity Analysis**

In this report, sensitivity analysis refers to an analysis where several models with different missing-data assumptions are jointly used for an incomplete longitudinal data set. As we argued in the introduction section, MPI is a powerful tool in separating the assessment on intermittent missingness from that on dropout missingness. In conducting investigations of dropout mechanisms, we have tools such as the D-K model to make further assessment on specific dropout mechanisms (e.g. ID vs. PDAR). However, such assessment is very sensitive to specifications of the repeated measures model and the dropout model. Therefore, we view the assessment of PDAR vs. ID as a part of sensitivity analysis instead of a standalone testing procedure. For details, see Chapter 4.
CHAPTER 3

MULTIPLE PARTIAL IMPUTATION

In this chapter, the multiple partial imputation (MPI) strategy will be introduced for handling missing data in longitudinal studies. In Chapter 2, we have seen that partial imputation of intermittent missing data offers a solution for assessing dropout mechanisms. In this chapter, we will see that it also provides a framework within which longitudinal modeling can be applied.

In section 3.1, we give a brief review on available missing-data methods for general incomplete data analysis, and explain why multiple imputation (MI) has become the most popular of these. Then, within the setting of longitudinal data analysis, the idea and algorithm for performing MPI are described in section 3.2. Section 3.3 discusses some computational considerations in application of MPI, including the EM algorithm, the Gibbs sampler and the associated convergence diagnoses. An imaginary data set and the smoking cessation data sets are used as examples for illustration purposes.

3.1 Taxonomy of Missing Data Methods

According to Little and Rubin (1987), missing-data methods can be categorized into three groups: (1) complete-case analysis where subjects with incomplete observations or responses are omitted, (2) analyzing as incomplete where all the observed part of the data are analyzed without deletion or imputation, and (3) imputation methods, where you fill in plausible values and then analyze the resulting complete data set by standard methods. For any of these approaches, either a model-based or empirical ad hoc procedure or algorithm can be used when handling missing values and making inferences.

To choose an appropriate approach, it is crucial to clarify the study goals beforehand. Are they related to parameter estimation (causal inference) or prediction? In clinical trials, the former is more often the situation. Correspondingly, Table 3.1 lists most of the common strategies that have been developed and applied in statistical practice for parameter estimation. For prediction, the paper by Sarle (1998) gives a nice compilation of the methods. It is often a difficult task to choose a proper strategy for a given incomplete dataset. Therefore, statistical analysis may be a post-hoc or context-dependent decision making process. As we discussed in Chapters 1 and 2, it is important to consider the missingness patterns and mechanism.
<table>
<thead>
<tr>
<th>Methods</th>
<th>Strength</th>
<th>Weakness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete Case Analysis</td>
<td>Analyze only fully observed subjects or cases. Discard all the incomplete cases.</td>
<td></td>
</tr>
<tr>
<td>List-wise Deletion</td>
<td>• Positive definite correlation matrix.</td>
<td>• Least powerful</td>
</tr>
<tr>
<td></td>
<td>• Simple.</td>
<td>• Requires MCAT</td>
</tr>
<tr>
<td>Weighting Methods</td>
<td>• Post. Definiteness.</td>
<td>• Takes time to create weights.</td>
</tr>
<tr>
<td></td>
<td>• Usually unbiased if MAT</td>
<td>• Many choices of weights</td>
</tr>
<tr>
<td>Analyzing as incomplete</td>
<td>Analyze all the observed data as if they were complete, but with no case-deletion or imputation.</td>
<td></td>
</tr>
<tr>
<td>Pair-wise deletion</td>
<td>• More powerful than list-wise deletion.</td>
<td>• Non-positive definiteness.</td>
</tr>
<tr>
<td>Structural Equation Modeling</td>
<td>• Powerful and elegant solution.</td>
<td>• Extremely tedious to program</td>
</tr>
<tr>
<td></td>
<td>• Uses all data</td>
<td>• Impractical for too much missing.</td>
</tr>
<tr>
<td></td>
<td>• Usually unbiased if MCAT</td>
<td></td>
</tr>
<tr>
<td>EM approach</td>
<td>• Incorporates stochastic error term in the imputation.</td>
<td>• No standard error.</td>
</tr>
<tr>
<td></td>
<td>• Usually unbiased if MAR</td>
<td>• Custom-programming.</td>
</tr>
<tr>
<td>Maximum Likelihood</td>
<td>• Accurate estimation.</td>
<td>• Difficult to implement.</td>
</tr>
<tr>
<td></td>
<td>• Model dependent.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Usually unbiased if MAR</td>
<td></td>
</tr>
<tr>
<td>GEE</td>
<td>• Robustness. Unbiased for regression coefficients.</td>
<td>• Requires MCAT.</td>
</tr>
<tr>
<td></td>
<td>• Usually unbiased if MAR</td>
<td>• Not likelihood-based.</td>
</tr>
<tr>
<td>Imputation Methods</td>
<td>Replace the missed values by some model-based imputation and then work on the imputed complete datasets.</td>
<td></td>
</tr>
<tr>
<td>Mean Substitution</td>
<td>• Simple to compute</td>
<td>• Artificially reduce variances and correlations.</td>
</tr>
<tr>
<td></td>
<td>• Usually unbiased if MCAT</td>
<td></td>
</tr>
<tr>
<td>Imputation by Regression</td>
<td>• Good prediction of the missing value</td>
<td>• Over-prediction.</td>
</tr>
<tr>
<td></td>
<td>• Usually unbiased if MCAT</td>
<td>• Multi-collinearity.</td>
</tr>
<tr>
<td>Subgroup Mean Substitution</td>
<td>• Reasonably simple.</td>
<td>• Gives inflated sample size that is biased, even if MAR</td>
</tr>
<tr>
<td></td>
<td>• Less biased than mean substitution</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Usually unbiased if MCAT</td>
<td></td>
</tr>
<tr>
<td>Multiple Imputation</td>
<td>• Powerful and elegant.</td>
<td>• Requires specialized program</td>
</tr>
<tr>
<td>Bayesian Methods</td>
<td>• Very flexible. Can describe the data in great detail.</td>
<td>• Takes long time to run;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Difficult custom-computation.</td>
</tr>
</tbody>
</table>

**Table 3.1** Common strategies for general missing data problems when the purpose of data analysis is parameter estimation

For most incomplete longitudinal data sets in substance abuse research, it is impossible to apply complete case analysis since the number of repeated measures is usually so large that even a small percentage of missing data on each measure would result in a large proportion of incomplete cases or subjects. Even if the
sample size is large enough or the number of repeated measure is small enough to afford such an analysis, it still has a serious limitation, requiring that missing data are MCAR.

At present many popular software packages (e.g., SAS, SPSS, and Splus) analyze incomplete longitudinal data using longitudinal models based on maximum likelihood or quasi-likelihood functions (e.g., GEE). When applying random effects models or Markov chain transition models the missing data should be at most MAR. Otherwise, biased estimation may be obtained. Similarly, GEE marginal models require even stronger assumption, that is, MCAR or DCAR within treatment conditions. There are some recent developments on longitudinal models that make it possible to handle nonignorable missing data. However, the application of them is not straightforward and few software packages have incorporated them.

A popular incomplete longitudinal data analysis strategy is to use imputation techniques to replace the missing data with some plausible values, which renders the data set complete and allows complete-data analysis. In substance abuse research, this method is more attractive based on the following facts.

1. Large amount of highly correlated repeated measures are capable of predicting each other, providing enough information for imputing missing data.
2. The job of making the imputations can be separated form the job of complete-data analysis, performed by different people at different times. Imputers may incorporate their understanding and available information (sometimes even confidential) on the missingness patterns and mechanisms into a model-based imputing algorithm.
3. Intermittent missing data and dropouts usually associate with different missingness mechanisms and can be handled differently when making imputations.

Sometimes, imputation of missing data is not explicitly seen in an incomplete data analysis, but it is implied. For example, traditional strategies for managing longitudinal missing data in the field of substance abuse research typically involve variations of the “missing as dirty” assumption (Ling et al. 1997): the data are missing due to active substance use or to consequences of substance use. When applying time-naïve models such as TES, the missing data points are actually imputed as “dirty” although no imputation is explicitly mentioned; see Chapter 2 for more information.

A final note is that except for the second factor listed above as supporting imputation methods, the other factors do not prefer imputation methods over analyzing as incomplete methods (see Table 3.1) or vice versa. Generally, if we have a model for imputation, we usually can transfer that model into an analysis using MLE or GEE without imputation. In these cases, the missingness mechanisms are well captured by the analytical models and it does not matter
much whether they are imputed or not. But it still matters which way is more convenient or efficient.

3.2 Multiple Imputation and Multiple Partial Imputation

3.2.1 Multiple Imputation

A serious limitation for standard imputation methods is that when treating the imputed values as if they are observed concrete values, the uncertainty naturally introduced by missingness becomes underestimated. In order to solve this problem, Rubin (1987) outlined multiple imputation, a Monte Carlo technique, in which the missing data are replaced by \( M > 1 \) sets of simulated values. Thus, \( M \) complete data sets are created and each of them is analyzed by standard complete-data methods. Then, the results are combined using a special rule to produce estimates and confidence intervals that incorporate missing-data uncertainty.

Data augmentation (Tanner and Wong 1987) is an MCMC technique that can be viewed as a variant of Gibbs sampling. Motivated by Little and Rubin (1987)’s missing data theory, Schafer (1997) developed a series of programs to implement data augmentation to accomplish multiple imputation.

Multiple imputation (Rubin 1987) is a Bayesian motivated method to simulate values of the parameter of interest \( \theta \) from the corresponding observed-data posterior distribution \( P(\theta | Y^{(o)}) = \int P(\theta | Y^{(o)}, Y^{(m)}) P(Y^{(m)} | Y^{(o)}) dY^{(m)} \), which can be approximated as

\[
P(\theta | Y^{(o)}) \cong \left[ \frac{1}{M} \sum_{l=1}^{M} P(\theta | Y^{(o)}, Y^{(m)}_{(l)}) \right]
\]

where \( Y^{(m)}_{(l)} \) are \( M \) independent imputations for \( Y^{(m)} \) which have been drawn from \( P(Y^{(m)} | Y^{(o)}) \), and \( M \) is a large enough number. Three steps are required to construct a MI process:

1. Create \( M \) independent plausible substitute values for \( Y^{(m)} \) based on some predictive distribution \( P(Y^{(m)} | Y^{(o)}) \);
2. Estimate \( \theta \) for each imputed complete data sets \( Y_{(i)} = (Y^{(o)}, Y^{(m)}_{(i)}) \);
3. Combine the \( M \) estimates into the final one as the result.

3.2.2 Multiple Partial Imputation

As we discussed in Chapters 1 and 2, in many substance abuse studies with longitudinal design, intermittent missing data usually account for a small proportion of the whole data matrix. Reasons underlying intermittent missingness
sometimes are not study-related, e.g., schedule conflicts, traffic problems, and bad weather conditions. Even if they were study-related, the usually large number of repeated measures (i.e., \( J \)) would convey enough information that could be used to predict on intermittent missingness. In other words, the associated missingness mechanism for intermittent missing data can be assumed to be or approximately be viewed as MCAR or MAR (Belin et al. 2000).

Nonetheless, the reasons underlying dropout are mostly related to the design of the study and dependent on the unobserved values due to dropout. Therefore, an assumption of MAR is usually far from the truth. In order to handle nonignorable dropouts using models such as Diggle and Kenward model or pattern mixture models, we have to differentiate them from intermittent close-to-random missing data.

In many substance abuse studies, the proportion of missing data points due to dropout in the whole data matrix is very large. It is suspect to make imputations on these dropouts, especially those that are far away from the observed data points, i.e., those in the right-lower corner in Figure 1.1.

Based on these considerations, we propose the idea of multiple partial imputation, which performs multiple imputations only on the intermittent missing data but leaves the dropouts as they are. This strategy has many other advantages. For example, by filling out the intermittent holes in the data set, the dropout mechanism can be assessed using the data-based evidence tests described in section 2.4. After partial imputation, the number of patterns of missingness drops significantly to a number that is affordable for a pattern-stratification analysis such as a pattern mixture model. For more discussion on MPI, please see Page 15.

### 3.2.3 A Bayesian Algorithm for MPI

Gibbs sampling is a Markov Chain Monte Carlo algorithm. In evaluating high dimensional probability density functions, Gibbs sampling iterates a series of simulations based on the fully conditional distributions, which is jointly equivalent to the simulating from the original high dimensional distribution. For an incomplete longitudinal normally distributed data, we perform the partial imputation on intermittent missing data by a Gibbs sampler proposed by Schafer (1997).

Suppose we partition the missing data \( Y_i^{(m)} \) further into \( Y_i^{(im)} = (Y_i^{(im)}, Y_i^{(dn)}) \), where \( Y_i^{(im)} \) represents intermittent missing data and \( Y_i^{(dn)} \) represents dropouts. Here, we treat \( Y_i^{(im)} \) as a vector of special parameters whose values are unknown and to be simulated. Assume that \( Y_i \sim N(\mu, \Sigma) = N(\mu(\beta), \Sigma(\alpha)) \), the Gibbs sampler for partial imputation is a process of iterations \( (t = 0,1,2,...) \), with each iteration containing two sub steps:
I-step: Given the current simulated values $\theta^{(t)} = (\beta^{(t)}, a^{(t)})$, for $i = 1, \ldots, N$, simulate $Y_i^{(im)}$ from

$$\begin{align*}
Y_i^{(im)(t+1)} &\sim P(Y_i^{(im)} | Y_{i}^{(o)}, \theta^{(t)}) \
\text{(3.2)}
\end{align*}$$

P-step: Conditioning on $Y^{(im)(t+1)} = \{Y_1^{(im)(t+1)}, \ldots, Y_N^{(im)(t+1)}\}$, draw a new value of $\theta$ from its posterior distribution given the current partially imputed data with monotone missingness patterns

$$\begin{align*}
\theta^{(t+1)} &\sim P(\theta | Y^{(o)}, Y^{(im)(t+1)}) \
\text{(3.3)}
\end{align*}$$

### Table 3.2 A Gibbs sampling imputation algorithm for MPI

In this sampler, the I-step involves a series of conditional distributions, which themselves are also normal distributions, from the theory on multivariate normal distribution. The P-step is carried out using the parameterization $\Phi = (\phi_1, \ldots, \phi_K)$ corresponding to $(Y^{(o)}, Y^{(im)})$, whose joint density can be factored as

$$\prod_{i=1}^{N} P(Y_{i1}, Y_{i2}, \ldots, Y_{id_i} | \Phi) = \prod_{i=1}^{N} P(Y_{i1} | \phi_1)P(Y_{i2} | Y_{i1}, \phi_2)\ldots P(Y_{id_i} | Y_{i1}, \ldots, Y_{id_{i-1}}, \phi_{d_{i-1}}).$$

Here $d_i \in \{2, \ldots, K\}$ indicates the dropout time point, and $d_i = K + 1$ means no dropout for subject $i$. There is a one-one map between $\theta$ and $\Phi$, and a numerical transformation using a sweeping approach is given by Schafer (1997).

In specifying the prior distributions for the parameters, we can apply the method of Boscardin and Weiss (2001), which specifies a prior distribution for $a$ through the use of an extra layer in Schafer’s hierarchical Bayesian model. That is, in addition to giving an Inverse-Wishart distribution to the covariance $(\Sigma_i)$ (Shafer 1997, Liu 1993), we will give a parametric form for the parameter of this inverse-Wishart distribution and then put a prior distribution on these parameters.

$$\begin{align*}
Y_i | \beta, \Sigma &\sim N(X_\beta, \Sigma_i) \text{ with } \Sigma_i = g(\Sigma) \
\Sigma^{-1} | a &\sim \text{Wishart}(\nu, \Omega(a)) \
p(\Sigma^-1 a) &\sim p(\beta) p(v) p(\nu) \
\text{(3.4-3.6)}
\end{align*}$$

### Table 3.3 Boscardin and Weiss’ hierarchical model proposed by

The idea in the above hierarchical model is to overdisperse the covariance matrix around parameterized structures. The amount of the overdispersion depends on the parameter $\nu$. Note that it is very difficult to precisely estimate the variance parameters using maximum likelihood techniques, yet in standard practice these variance estimates are used as if they were infinitely precise. By using a Bayesian approach, many difficulties associated with ML approaches can be solved.

By running parallel Gibbs samplers from $M$ distinct starting points, we can get $M$ Markov chains. Then, imputed data sets can be created from each chain after a long enough burn-in period. Or we may generate the $M$ imputed sets from a single
chain with long enough intervals between any two consecutive imputations. In
order to yield valid inference, a proper multiple imputation must possess certain
properties discussed by Rubin (1987) and Schafer (1997), one of which is that it
must independently realize \( P(Y^{(o)}, Y^{(m)}) \). Another approach for applying the above
Gibbs sampler is to merge the data analysis model into this Gibbs structure, so
that the posterior distribution of the parameters of interest will be evaluated
directly.

### 3.3 Consolidating Multiple Analysis Results using Rubin’s Rule

As mentioned earlier, after performing multiple complete-data analyses, we need
a way to combine these analysis results to make a final inferential statement.
Using a scalar parameter as an example, the final expectation and variance of \( \theta \)
are calculated based on the following well-known statistical results:

\[
E(\theta | Y^{(o)}) = E\left[E(\theta | Y^{(o)}, Y^{(m)}) | Y^{(o)}\right] \tag{3.7}
\]

and

\[
Var(\theta | Y^{(o)}) = E[Var(\theta | Y^{(o)}, Y^{(m)}) | Y^{(o)}] + Var[E(\theta | Y^{(o)}, Y^{(m)}) | Y^{(o)}] \tag{3.8}
\]

For a univariate \( \theta \), let \( \hat{\theta}^{(i)} \) and \( U^{(i)} \) designate its point and variance estimates
using the \( l^{\text{th}} \) set of imputed data (i.e. \( Y^{(l)} = (Y^{(o)}, Y^{(m)}) \)). The final point estimate
for \( \theta \) is simply the average of the complete-data point estimates,

\[
\bar{\theta} = \frac{1}{M} \sum_{l=1}^{M} \hat{\theta}^{(i)}. \tag{3.9}
\]

The variance estimate associated with \( \bar{\theta} \) is

\[
T = \bar{U} + (1 + M^{-1})B, \tag{3.10}
\]

which is an estimate that incorporates the uncertainty due to missing data, where

\[
\bar{U} = \frac{1}{m} \sum_{l=1}^{m} U^{(i)} \text{ is the average within-imputation variance and } B = \frac{1}{m-1} \sum (\hat{\theta}^{(i)} - \bar{\theta})^2.
\]

is the between-imputation variance. A 100(1-\( \alpha \))% interval estimates for \( \theta \) can be
obtained as

\[
\bar{\theta} \pm t_{\nu,\alpha/2} \sqrt{T} \tag{3.11}
\]

where \( \nu = (M-1)(1+1/r)^2 \) is the degrees of freedom of the \( t \) distribution and

\[
r = (1+1/M)B / \bar{U} \tag{3.12}
\]

is referred as the relative increase in variance due to missingness. For the case of
multi-dimensional \( \theta \), please refer to Rubin (1987, pp. 76-78).

Adopting the information defined by Fisher (1935), Rubin (1987) proposed a
measure of the fraction of missing information about \( \theta \), given as

\[
r + 2/(\nu + 3) \tag{3.13}
\]

Table 3.4 offers a guideline for how many imputations are needed in performing
multiple imputation, a question evaluated by the percent efficiency achieved for
various rates of missing information and values of \( m \). Unless the fraction of
missing information is extremely high, the efficiency gained by taking more than 5-10 imputations is minimal.

<table>
<thead>
<tr>
<th>$M$</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>97</td>
<td>91</td>
<td>86</td>
<td>81</td>
<td>77</td>
</tr>
<tr>
<td>5</td>
<td>98</td>
<td>94</td>
<td>91</td>
<td>88</td>
<td>85</td>
</tr>
<tr>
<td>10</td>
<td>99</td>
<td>97</td>
<td>95</td>
<td>93</td>
<td>92</td>
</tr>
<tr>
<td>20</td>
<td>100</td>
<td>99</td>
<td>98</td>
<td>97</td>
<td>96</td>
</tr>
</tbody>
</table>

**Table 3.4** Percent efficiency achieved for various combination of fraction of missing information and number of imputations (i.e., $M$). (From Rubin (1987))

### 3.4 Other Computational Considerations in Applying MPI

When creating partial imputations in applying MPI, a Gibbs sampler (see Table 3.2) is used. In this section, several computational issues are considered. First, the definition of MCMC and Gibbs sampler will be given. Then we consider how to specify a starting point using the EM-algorithm for the parameter to be simulated in the Gibbs sampler for MPI. Another issue to be discussed is convergence diagnostics. Finally, an imaginary data set is used as an example.

#### 3.4.1 Markov Chain Monte Carlo (MCMC) and Gibbs Sampler

MCMC methodology, as a powerful computational tool, became popular in the statistical community only recently after the publications of Gelfand and Smith (1990) and Gelfand et al. (1990), although it was introduced by Metropolis et al. (1953) almost half a century ago.

A Markov chain is a stochastic process \{\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(t)}\}\} that satisfies the properties:

1. \( f(\theta^{(t+1)} | \theta^{(t)}, \ldots, \theta^{(1)}) = f(\theta^{(t+1)} | \theta^{(t)}) \). That is, the distribution of \( \theta \) in time \( t+1 \) given all the preceding \( \theta \) (i.e. \( \theta^{(1)}, \ldots, \theta^{(t)} \)) depends only on \( \theta^{(t)} \).

2. \( f(\theta^{(t+1)} | \theta^{(t)}) \) is also independent of time \( t \).

If such a Markov chain is irreducible, aperiodic and positive recurrent, then the distribution of \( \theta^{(t)} \) tends to its equilibrium distribution, which is independent of the initial \( \theta^{(0)} \), as \( t \to \infty \). This definition is from Gilks et al. (1996).

MCMC is the method of Monte Carlo based on the above Markov chains. When applying MCMC to generate a sample from \( f(\theta | y) \), we can construct a Markov chain with this distribution (i.e. \( f(\theta | y) \)) as the equilibrium distribution. Starting this Markov chain from a selected initial value \( \theta^{(0)} \), after a burn-in period with \( t_0 \) iterations, \( \{\theta^{(t_0+1)}, \theta^{(t_0+2)}, \ldots, \theta^{(t_0+t)}\} \) are generated as desired samples from \( f(\theta | y) \). The burn-in period is used to ascertain that the equilibrium distribution is reached or the chain is converged. The convergence can be monitored or checked by
various methods. Of course, given the Markov structure, these samples are correlated. One way to obtain approximately independent samples is to take draws separated by considerable lags, so as to ensure that the autocorrelation between draws remains negligible.

In the construction of the above Markov chain, \( f(\theta^{(r+1)} | \theta^{(r)}) \) should ideally be easy to generate. Different approaches to realize this transition function determine various MCMC algorithms. Two popular algorithms are the Metropolis-Hastings algorithm (Metropolis et al. 1953, Hastings 1970) and the Gibbs sampler (Geman and Geman 1984).

**Gibbs Sampler**

The Gibbs sampler is an algorithm to evaluate a high dimensional posterior distribution \( f(\theta | y) \) by iteratively evaluating one component at a time of the parameter vector (i.e., \( \theta \)) from a sequence of conditional posterior distributions. Starting a Markov chain from state \( \theta^{(0)} \), suppose the current state becomes \( \theta^{(r)} \), the next iterate is achieved through the following sub-steps:

\[
\begin{align*}
\text{sample } \theta^{(r+1)}_1 & \text{ from } f(\theta_1 | \theta_2^{(r)}, \theta_3^{(r)}, ..., \theta_p^{(r)}, y) \\
\text{sample } \theta^{(r+1)}_2 & \text{ from } f(\theta_2 | \theta_1^{(r+1)}, \theta_3^{(r)}, ..., \theta_p^{(r)}, y) \\
\text{sample } \theta^{(r+1)}_3 & \text{ from } f(\theta_3 | \theta_1^{(r+1)}, \theta_2^{(r+1)}, ..., \theta_p^{(r)}, y) \\
& \vdots \\
\text{sample } \theta^{(r+1)}_p & \text{ from } f(\theta_p | \theta_1^{(r+1)}, \theta_2^{(r+1)}, ..., \theta_{p-1}^{(r+1)}, y)
\end{align*}
\]

where \( p \) is the number of components of \( \theta \), i.e. \( \theta = (\theta_1, \theta_2, ..., \theta_p) \). Usually \( f(\theta_j | \theta_1^{(r+1)}, ..., \theta_{j-1}^{(r+1)}, \theta_{j+1}^{(r+1)}, ..., \theta_p^{(r)}, y) \) has a much simpler form than \( f(\theta | y) \), thus simulating components of \( \theta \) one by one is usually much easier that simulating them simultaneously. For more details on the Gibbs sampler, see Casella and George (1992) and Gilks et al. (1996).

For Gaussian distributed longitudinal data, the algorithm described in Table 3.2 is a Gibbs sampler, where \( \theta = (\mu, \Sigma, Y^{(m)}) \) and \( Y^{(m)} \) can be viewed as a special type of parameter whose values are uncertain. When initiating this Gibbs sampler, we need a starting value for \( \theta \). As Schafer (1997) suggested, a maximum likelihood estimate \( (\hat{\mu}, \hat{\Sigma}) \) given by the EM algorithm offers a good choice.

### 3.4.2 EM-Algorithm and Starting Points

Taking advantage of the interdependence between missing data \( Y^{(m)} \) and parameters \( \theta \), an intuitive way for estimating \( \theta \) is an iterative procedure, which, starting from an initial value for \( \theta \), repeats the following two steps until \( \theta \)
converges: (1) Impute the missing data \( Y^{(m)} \) based on \( \theta \) and the observed data \( Y^{(o)} \); (2) re-estimate \( \theta \) based on \((Y^{(m)}, Y^{(o)})\).

For incomplete problems, the distribution of the complete data \( Y \) can be factored as \( P(Y|\theta) = P(Y^{(o)}|\theta)P(Y^{(m)}|Y^{(o)},\theta) \). Correspondingly, the log likelihood function for \( \theta \) can be written as

\[
l(\theta | Y) = l(\theta | Y^{(o)}) + \log(P(Y^{(m)} | Y^{(o)},\theta)) \tag{3.14}
\]

where \( l(\theta | Y^{(o)}) \) is the observed-data log likelihood function (also see 2.4), and \( P(Y^{(m)} | Y^{(o)},\theta) \) is called the predictive distribution of the missing data. Since \( Y^{(m)} \) is unknown, we can integrate them out on both side of equation (3.1) with regard to \( P(Y^{(m)} | Y^{(o)},\theta^{(i)}) \), that is,

\[
Q(\theta | \theta^{(i)}) = l(\theta | Y^{(o)}) + H(\theta | \theta^{(i)}) \tag{3.15}
\]

where

\[
Q(\theta | \theta^{(i)}) = \int l(\theta | Y)P(Y^{(m)} | Y^{(o)},\theta^{(i)})dY^{(m)}, \tag{3.16}
\]

and

\[
H(\theta | \theta^{(i)}) = \int \log(P(Y^{(m)} | Y^{(o)},\theta))P(Y^{(m)} | Y^{(o)},\theta^{(i)})dY^{(m)}. \tag{3.17}
\]

Let \( \theta^{(r(i))} \) represent the value of \( \theta \) that maximizes \( Q(\theta | \theta^{(i)}) \), then, as Dempster, Laird, and Rubin (1977) proved,

\[
l(\theta^{(r(i))} | Y^{(o)}) \geq l(\theta^{(i)} | Y^{(o)}), \tag{3.18}
\]

which means that the observed-data log likelihood function of \( \theta^{(r(i))} \) is larger or equal to that of \( \theta^{(i)} \). In other words, an increase in \( Q(\theta | \theta^{(i)}) \) always accompany by an increase in \( l(\theta | Y^{(o)}) \).

In the Expectation-Maximization or EM algorithm, each iteration consists of the following two steps:

1. Expectation step or E-step: calculate \( Q(\theta | \theta^{(i)}) \) using (3.16);

2. Maximization step or M-step: find \( \theta^{(r(i))} \) by maximizing \( Q(\theta | \theta^{(i)}) \).

Repeating the above E- and M-steps with a starting value \( \theta^{(0)} \), we obtain a sequence of iterates \( \theta^{(0)}, \ldots, \theta^{(r)} \), which converges to a stationary point of the observed-data log likelihood under regularity conditions given by Dempster, Laird and Rubin (1977) and Wu (1983). In many well-behaved problems, this stationary point is the global maximum of \( l(\theta | Y^{(o)}) \), i.e., the maximum likelihood estimate of \( \theta \).

For a multivariate normally distributed longitudinal data set, the formulations in E-step and M-step have simple forms (Dempster, Laird, and Rubin 1977). When creating multiple partial imputations using a single chain of Gibbs sampler, the stationary point \((\hat{\mu}, \hat{\Sigma})\) reached by the EM algorithm can be used as the starting point to trigger the I- and P-steps (see Table 3.2).
3.4.3 Convergence Diagnosis

Another critical question for applying MPI Gibbs sampler is when it is safe to stop the sampling process and believe that the equilibrium distribution has been reached. Analytical approaches like theoretical convergence bounds have a promising future but are currently not able to be used in practice. Most Bayesian practitioners check convergence by applying some diagnostic tools on the output from the samplers.

In order to make valid inference, multiple partial imputations for the intermittent missing data must be created properly in the sense that they must be independently drawn from the stationary distribution. For the set of simulations \( \{(\theta^{(t)}, Y^{(im)}_{(t)}): t = 0,1,2,...\} \) to be distributed from \( P(\theta, Y^{(im)} | Y^{(o)}) \), it is sufficient for the distribution of \( \{\theta^{(t)}: t = 0,1,2,...\} \) to have converged to \( P(\theta | Y^{(im)}) \). This is because \( \theta^{(t)} \sim P(\theta | Y^{(o)}) \) implies that \( (\theta^{(t)}, Y^{(im)}_{(t)}) \sim P(\theta, Y^{(im)} | Y^{(o)}) \) for all \( t' > t \). Therefore, we can determine the convergence of \( Y^{(im)}_{(t)} \) by monitoring the convergence of \( \theta^{(t)} \) which usually has a much smaller dimensionality.

Cowles and Carlin (1996) and Mengerson et al. (1999) give fairly extensive comparative reviews on available diagnostic tools. Most of them are popularly used in practice. Two free software packages with some general-purpose diagnostic tools are now available in the Bayesian community. CODA (Convergence Diagnosis and Output Analysis Software) by Best, Cowles, and Vines (1995), is a supplemental component of the widely-used MCMC software BUGS. The other is BOA (Bayesian Output Analysis) was developed by Smith (2001) as an extension of CODA. Since these software packages are freely available, we decided not to implement them in MPI 2.0. For diagnostic purpose, MPI 2.0 saves the simulated parameters and then exports them for further investigation using CODA or BOA.

Convergence Diagnostic Tools for Individual Chains

It is more convenient to simulate a long single chain of Markov sequence when creating multiple partial imputations. For an individual component (e.g., \( \mu_{t} \)) or a scalar function of the parameters of interest (e.g., \( h(\mu) \)), time-series plots and autocorrelation plots are convenient but simple tools to assess their convergence. In most cases, we need more advanced tools that offer more accurate assessment. The most popularly used ones are those proposed by Raftery and Lewis (1992), by Geweke(1992) and by Heidelberger and Welch (1983).

The Raftery and Lewis convergence diagnostic not only tests for convergence of the chain to its stationary distribution, but also provides a way of bounding the
variance of estimates of quantiles of functions of parameters. When convergence of the mean of some function of the sampled parameters is of interest, Geweke convergence diagnostic compares the mean of the sampled values in the first window containing the first \( n_1 \) iterations to the mean of the sampled values in the last window containing the last \( n_2 \) iterations. A \( Z \) statistic is calculated as the difference between the two means divided by the asymptotic standard error determined by spectral density estimation. The Heidelberger and Welch convergence diagnostic is based on Brownian bridge theory and uses the Cramer-Mises statistic.

**Convergence Diagnostic Tools for Multiple Parallel Chains**

Sometimes, we may simultaneously run several short chains to create multiple partial imputations. In many cases, the convergence seems to be achieved quickly even if a single chain is used. Nevertheless, when multiple parallel chains are used, starting from over-dispersed points, we can detect slow convergence by observing differences across the chains.

When running multiple chains, we should set starting points that are separated as far as possible from each other in the parameter space. Several methods for generating starting values for the MCMC samplers have been proposed (Gelman and Rubin 1992, Applegate et. al. 1990, Jennison 1993). We suggest a method using a preliminary single chain. After running a single chain from a starting point, e.g., \((\hat{\mu}, \hat{\Sigma})\) estimated by EM, we collect all the simulated values for \((\mu, \Sigma)\) from which over-dispersed points are chosen.

Based on normal theory approximations to exact Bayesian posterior inference, Gelman and Rubin (1992) proposed the potential scale reduction factor (PSRF) to test whether multiple chains are converged for each scalar quantity of interest. Suppose the number of iterations within each chain is \(2N\), then, the PSRF is defined as

\[
\sqrt{R} = \sqrt{\frac{N - 1}{N} + \frac{M + 1}{MN} \frac{B}{W}} \frac{df}{df - 2}
\]

(3.19)

where \(B\) is the variance between the means from the \(M\) parallel chains, \(W\) is the average of the \(M\) within-chain variances, and \(df\) is the degree of the approximating \(t\) density. Note that this statistic employs a similar idea to that used adopted by Rubin for multiple imputation. Other diagnostic tools for multiple chains include the corrected scale reduction factor (CSRF) and the multivariate potential scale reduction factor (MPSRF) proposed by Brooks and Gelman (1998).

**Missing Information and Worst Linear Function**

An important factor that affects the convergence speed is the missing information (using Fisher’s definition). According to the missing information principle of
Orchard and Woodbury (1972), the complete information is equal to the observed information plus the missing information. In the context of MI, Rubin gives the formula of fraction of missing information (FMI); see (3.8). When obtaining the maximum likelihood estimate of $\theta$ by EM algorithm, Dempster, Laird and Rubin (1977) found that the convergence speed of the EM was determined by the FMI. When $\theta$ is a vector of dimensionality higher than one, FMI is not a scalar but instead a matrix. We call the eigenvector corresponding to the largest eigenvalue of this matrix the maximum missing information fraction eigenvector (MMIFE), which governs the rate of convergence of the EM algorithm. Similarly to EM, the Gibbs sampler for missing data imputation will also be affected by the missing information with the consideration that the convergence rate of $\theta^{(t)}$ is slowest in the direction determined by MMIFE.

Many convergence diagnostic tools are applied to check convergence of each component or scalar functions of the parameters of interest. When $\theta^{(t)}$ has a large dimensionality, many scalar functions exist. The relevant question is which one to choose in order to check the overall convergence. If we identify one whose marginal distribution converges slowest, then the convergence of this worst function would guarantee a global convergence. By restricting attention to the linear combination of the components of $\theta^{(t)}$, in the form of $V^T\theta^{(t)}$, Schafer proposes the so-called worst linear function (WLF) where vector $V$ is set as MMIFE. It has the largest rate of missing information, hence the slowest convergence speed. An estimate of MMIFE can be obtained from the convergence behavior of EM with the last several steps approximately linear for $\hat{\theta} - \theta^{(t)}$ ($t = ..., T - 1, T$). An approximation of MMIFE, therefore, is

$$V = \hat{\theta} - \theta^{(T-t)},$$

where $\hat{\theta} = \theta^{(T)}$ is the converged point of EM.

### 3.4.4 Illustration Using an Imaginary Data Set

Table 3.5 gives an incomplete bivariate data set with twelve sample units originally reported by Murray (1977) in his comment to the seminal paper on EM algorithm by Dempster, Laird and Rubin’s (1977). We can view this data set as a longitudinal one with at most two repeated measures observed for each of twelve cocaine dependents. For the following discussion, we assume that the complete data are i.i.d. $\begin{pmatrix} Y \\ X \end{pmatrix} \sim N\left( \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_X^2 \end{pmatrix} \right)$ and the missingness mechanism is ignorable.

<table>
<thead>
<tr>
<th>Y</th>
<th>1</th>
<th>1</th>
<th>-1</th>
<th>-1</th>
<th>2</th>
<th>2</th>
<th>-2</th>
<th>-2</th>
<th>?</th>
<th>?</th>
<th>?</th>
<th>?</th>
</tr>
</thead>
</table>

Table 3.5 An incomplete bivariate data set
For this data set, we first run the EM algorithm with three different starting points for \((\mu_Y, \mu_X, \sigma_Y^2, \sigma_{XY}^2, \sigma_X^2)\): \((0, 0, 1, 1, -0.2), (0, 0, 1, 1, 0), (0, 0, 1, 1, 0.2)\). The starting values differ only in \(\sigma_{XY}^2\), which is mostly affected by missing data. From Table 3.6, we see that for the starting point with \(\sigma_{XY}^2 = 0\), the EM algorithm takes 16 iterations to converge but 118 iterations are required if starting from the other points. Also notice that the three converged points for \((\mu_Y, \mu_X, \sigma_Y^2, \sigma_{XY}^2, \sigma_X^2)\) are different, which means that multiple modes or saddle-points exist. From a numerical derivation, we can verify that two of them, \((0, 0, 2.67, 2.67, -1.33)\) and \((0, 0, 2.67, 2.67, 1.33)\), are local modes and one, \((0, 0, 2.5, 2.5, 0)\), is a saddle point (see Figure 3.3c). The estimated approximate MMIFE can be used to create the worst linear functions, which can be used to monitor the convergence of Gibbs sampler for Multiple Imputation.

<table>
<thead>
<tr>
<th>(\sigma_{XY}^2)</th>
<th>#Iterations</th>
<th>Converged Points</th>
<th>Approximate MMIFE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-0.2)</td>
<td>118</td>
<td>((0, 0, 2.67, 2.67, -1.33))</td>
<td>((0, 0, 0.023, -0.099, 0.023))</td>
</tr>
<tr>
<td>(0)</td>
<td>16</td>
<td>((0, 0, 2.5, 2.5, 0))</td>
<td>((0, 0, 0.07, 0, 0.07))</td>
</tr>
<tr>
<td>(0.2)</td>
<td>118</td>
<td>((0, 0, 2.67, 2.67, 1.33))</td>
<td>((0, 0, 0.023, 0.099, 0.023))</td>
</tr>
</tbody>
</table>

**Table 3.6** Iterations required by the EM algorithm and the approximate eigenvector corresponding to the maximum MIF

We next run a Gibbs sampler to create multiple imputations for the missing data. Note that we do not differentiate intermittent missing data from dropouts in this example and Schafer’s data augmentation algorithm (Schafer 1997) with non-informative prior distributions is used here. We start the Gibbs sampler from the three converged points. Thus three parallel chains are simulated, each with 100 iterations. When running an EM algorithm, once a converged local maximum point is reached, the iterative process stops. In running the Gibbs sampler to create multiple imputations, we hope that the sampler will not get “stuck” at any local modes, i.e., the corresponding Markov chain should be irreducible. Figure 3.1 gives the trace plots for all five parameters simulated with 100 iterations ((MeanY, MeanX, VarY, CovYX, VarX) = \((\mu_Y, \mu_X, \sigma_Y^2, \sigma_{XY}^2, \sigma_X^2)\)).

From the trace plots, we see that for each parameter the three parallel chains become identical within 40 iterations but there is evidence that each individual chain does not converge. Applying Geweke diagnostic to the simulated values of the parameter CovYX, we have Z-score of \(-3.132\) (p-value=0.0017) for the first chain and Z-score \(-1.989\) (p-value=0.046) for the second chain. By applying Heidelberger & Welch diagnostic, the test fails to pass for the third chain. By calculating the autocorrelations we can find that the Lag50 autocorrelations are still significantly non-zero for CovYX in all three parallel chains. However, if we only apply Brooks, Gelman & Rubin diagnostics, we would get a completely different conclusion, i.e., no evidence to reject the hypothesis of convergence. A
possible reason for this conflict is that the three starting points are not over-dispersed.

![Figure 3.1](image1.png)

**Figure 3.1** Trace plots of the parameters \((\mu_Y, \mu_X, \sigma_Y^2, \sigma_{XY}^2, \sigma_X^2)\) with 100 iterations

If we try to simulate more chains starting from a wide range of values, we would still observe that the behaviors of the chains quickly become identical. These trials give us further evidence that the imputation Gibbs sampler might have mixed after approximately 50 iterations, although individual-chain diagnostics suggest contrary evidence. In order to confirm this, we run the Gibbs sampler once again but with 500 iterations this time. Figure 3.2 gives the trace plots for the three parallel chains each with 500 iterations. This time, it is apparent that the chains have converged for all the parameters except CovYX. For the first chain starting from \((0, 0, 2.67, 2.67, -1.33)\), the autocorrelation plot shows that dependence exists between two simulated samples that can be as far as 25 iterations away from each other. The Gewek diagnostic plot depicts a picture where the first chain is on the edge between convergence and non-convergence. This phenomenon should have been anticipated by an initial review of the original data set where the missingness pattern means the data convey little information for CovYX.

Continuing this type of procedure, let us increase the number of iterations as large as 5000. Figure 3.3 shows the trace plot, Geweke diagnostic plot, and the smoothed marginal density function of the 5000 simulated values for \(\sigma_{XY}^2\) (i.e. CovYX) for one chain starting from \((0, 0, 2.67, 2.67, -1.33)\). All plots and other diagnostic tools also indicate that convergence is achieved. For other chains starting from other points, we can make similar conclusions and obtain very similar plots.
From the above exploration process, it may be reasonable to conclude that there is little evidence to reject the hypothesis that the imputation Gibbs sampler converges after 5000 iterations. When we create multiple imputations using a single chain, we can set the imputation interval by any number larger than 50 because the autocorrelation approximately decreases to zero after 50 lags for all components of $\theta_{IMP}$. Another important finding is that, by applying the Raftery and Lewis convergence diagnostic tool to the chains with 5000 iterations, we might thin the Markov chain with a thinning interval of 3.

**Figure 3.2** Trace plots of the parameters $(\mu_Y, \mu_X, \sigma^2_Y, \sigma^2_{XY}, \sigma^2_X)$ with 500 iterations

**Figure 3.3a** Trace plot for $\sigma^2_{XY}$ in the first chain with 5000 iterations

**Figure 3.3b** Geweke diagnostic for $\sigma^2_{XY}$ in the first chain with 5000 iterations
In previous analyses, we did not use WLFs for diagnostic. For this small data set, with only five parameters involved and $\sigma^2_{XY}$ converges slowly relative to the others, monitoring of $\sigma^2_{XY}$ is approximately equivalent to monitoring any WLFs.

### 3.5. Application of MPI to data from the Smoking Cessation Study

For the smoking cessation study, from Table 1.2 we see very few intermittent missing values for self-reported number of cigarettes. Therefore, the fraction of missing information for the intermittent missing data is low and the convergence of the MPI Gibbs samplers should be less problematic. In this section, we use the carbon monoxide level data as illustrations of performing MPI in MPI 2.0. The following MPI manipulations were performed on a 12-measure subset of the original 36-measure data of repeated carbon monoxide levels; see the User’s Guide for details.

#### 3.5.1 Maximum Likelihood Estimation Using EM-Algorithm

In MPI 2.0, before creating multiple partial imputations, the EM-algorithm is run to obtain the maximum likelihood estimate ($\hat{\mu}, \hat{\Sigma}$). This estimate can be used as a starting value to trigger a MPI Gibbs sampler. After specifying the maximum number of iterations as 1000 and the convergence criteria as $0.0000001$, the EM-algorithm was run in MPI 2.0. On an IBM PC with 1.72 GHZ CPU and 256MB memory, within 3 seconds, the procedure converged for carbon monoxide levels in CMs and No-CMs groups. The converged values for the mean vector and the covariance matrix, i.e., $(\hat{\mu}, \hat{\Sigma})$, are shown graphically and numerically.

Figure 3.4 depicts the EM estimates of the mean vectors (i.e., $\hat{\mu}$) for participants who received contingency management (CMs=Group 2) and those did not received contingency management (No-CMs=Group 1). We can compare this plot (Plot 2.1 of Figure 3.4) with the marginal means shown by Plot 1.2 of Figure 2.3 in the User’s Guide, to see how different they are from each other. The marginal means were calculated using the original observed values without EM. The closer the EM estimated means to the marginal means, the smaller the fraction of missing information caused by intermittent missingness.
Figure 3.4 EM estimated mean vectors (i.e., $\hat{\mu}$) for repeated CO-levels within CMs group (Group 2) and No-CMs group (Group 2)

Figure 3.5 Correlation structures estimated by the EM-algorithm for carbon monoxide level data (Group 1 = No-CMs, Group 2=CMs)

<table>
<thead>
<tr>
<th>-- Group 1=No-CMs --</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
</tr>
<tr>
<td>Correlation</td>
</tr>
<tr>
<td>Y1</td>
</tr>
<tr>
<td>Y2</td>
</tr>
<tr>
<td>Y3</td>
</tr>
<tr>
<td>Y4</td>
</tr>
<tr>
<td>Y5</td>
</tr>
<tr>
<td>Y6</td>
</tr>
<tr>
<td>Y7</td>
</tr>
<tr>
<td>Y8</td>
</tr>
<tr>
<td>Y9</td>
</tr>
<tr>
<td>Y10</td>
</tr>
<tr>
<td>Y11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>-- Group 2=CMs--</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
</tr>
<tr>
<td>Means</td>
</tr>
<tr>
<td>Correlation</td>
</tr>
<tr>
<td>Y1</td>
</tr>
<tr>
<td>Y2</td>
</tr>
<tr>
<td>Y3</td>
</tr>
<tr>
<td>Y4</td>
</tr>
<tr>
<td>----</td>
</tr>
<tr>
<td>Y5</td>
</tr>
<tr>
<td>Y6</td>
</tr>
<tr>
<td>Y7</td>
</tr>
<tr>
<td>Y8</td>
</tr>
<tr>
<td>Y9</td>
</tr>
<tr>
<td>Y10</td>
</tr>
<tr>
<td>Y11</td>
</tr>
</tbody>
</table>

**Table 3.7** Means and correlation coefficients for carbon monoxide data estimated by the EM algorithm

Figure 3.5 graphically depicts the correlation structure (i.e., standardized $\hat{\Sigma}$) estimated from the EM algorithm. Within each 12 by 12 bar matrix, each bar represents the correlation coefficients ($0 \leq \rho_{ij} \leq 1$) between any pair of repeated carbon monoxide levels. These correlation structures provide useful evidence for covariance structure specification in fitting longitudinal models such as marginal models or mixed models. For the No-CMs group (Group 1), we see that the farther apart the two measures are from each other in time, the smaller the correlation coefficient between them. But such a pattern does not apply to the correlation structure within CMs groups.

Table 3.7 lists the mean values and correlation coefficients corresponding to the plots in Figures 3.4-3.6. Note that in MPI 2.0, users have the option to see the estimated covariance matrix (i.e., $\hat{\Sigma}$).

### 3.5.2 Creating Multiple Partial Imputations for Carbon Monoxide Levels

We ran a single chain to create multiple partial imputations for the carbon monoxide data. After a burn-in period of 1000 iterations, we created four imputations using the Gibbs Sampler described in Table 3.2 with an imputation interval of 500 iterations. On the same computer, this Gibbs sampler with altogether 3000 iterations of simulation took 5 seconds. The seed for the random generator was saved so that the simulation process could be repeated. The simulated parameters during burn-in period were saved to files on the hard disk. Figure 3.6 depicts the *track plots* (or *time-series plots*) of WLFs for all the parameters, $(\mu, \Sigma)$, of the data within each collapsed group (No-CMs or CMs). Visually, we observed strong evidence in support of convergence from the two time-series for WLFs, especially for WLFs in Group 1 (i.e., No-CMs). This is consistent with the results of testing on MCAR for intermittent missing CO-levels in Chapter 1, which says that MCAR holds for people receiving no contingency management (i.e., No-CMs).
Using a procedure of convergence diagnostic similar to the one for the imaginary data in Table 3.5, we determined the burn-in length. For the 1000 burn-in values of WLFs depicted in Figure 3.6, the lag-50 autocorrelation functions are close to 0 for both WLF1 and WLF2 (see Figure 3.7). Using Geweke’s diagnostic, Figure 3.8 shows the Z-scores, from which we observe moderate evidence for rejecting the hypothesis of convergence of the Gibbs samplers measured by WLFs. Five out of ten and six out of ten Z-scores are scattered away from the null interval, (-1.98, 1.98), for WLF1 and WLF2. Both WLF1 and WLF2 with 1000 simulations passed the Heidleberger and Welch stationarity tests, which means that the convergence should be accepted within 1000 iterations. Another useful tool for convergence diagnostic is the running means, which are shown in Figure 3.9, which suggests that the mixture or convergence may start after 200 iterations.
Using this Gibbs sampler, 4 partially imputed data sets are created for CO-levels, all saved in ASCII format on the hard disk. The next step is to apply appropriate modeling strategies to these imputed data sets according to our investigation on missingness patterns and mechanisms performed in Chapters 1 and 2. This will be the topic of Chapter 4. The following discussion on MPI-inference is for demonstration purposes, and should not be viewed as the final analysis for the smoking cessation study. For a better sense of a more complete data analysis, please refer to the next chapter.

3.5.3 Making MPI-Inference Using Rubin’s Rule

Here, we use the 12-measure CO-level data set as an example to show how to make MPI-inferences when each of the partially imputed data sets is analyzed using a random-effects model with random intercept and slope. From Figure 3.4, we saw that the two mean CO profiles declined differentially during the first two weeks and thereafter both seemed to level off. Therefore, we fit the following piecewise-linear random effects model to each partially imputed data set,

\[ Y_{ij} = \beta_0 + \beta_1 T_i + \beta_2 T_i CM + \beta_3 T_i CM + \beta_4 T_i + \beta_5 T_i CM + \gamma_{0i} + \gamma_{1i} T_i CM + \epsilon_{ij}, \]

where predictors are defined as:

\[ T = \begin{cases} 
0 & \text{if } Time \leq 2\text{wk} \\
1 & \text{if } Time > 2\text{wk}
\end{cases}, \quad T_i = \begin{cases} 
0 & \text{if } Time \leq 2\text{wk} \\
1 & \text{if } Time > 2\text{wk}
\end{cases}, \]

\[ T_s = \begin{cases} 
0 & \text{if } Time \leq 2\text{wk} \\
1 & \text{if } Time > 2\text{wk}
\end{cases}, \quad \text{and } CM = \begin{cases} 
0 & \text{if participant is in No-CMs group} \\
1 & \text{if participant is in CMs group}
\end{cases}. \]

In this random-effects model, coefficients \( \gamma_{0i} \) and \( \gamma_{1i} \) respectively represent random intercept effect and random slope effect.

To test treatment effectiveness of the interventions involving contingency management (i.e., between CMs and No-CMs), we fitted the model on each of the four partially imputed data set created in section to test whether \( \beta_2 \) is significantly different from zero, i.e., \( \beta_2 = 0 \). After fitting the model to the four imputed data sets, the four estimates of \( \beta_2 \) in Table 3.8 were obtained.

<table>
<thead>
<tr>
<th>Imputation #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point Estimate (( \hat{\beta}_2 ))</td>
<td>2.3755</td>
<td>2.2867</td>
<td>2.249</td>
<td>2.4035</td>
</tr>
<tr>
<td>Variance (Var(( \hat{\beta}_2 )))</td>
<td>0.8859</td>
<td>0.8968</td>
<td>0.9129</td>
<td>0.7324</td>
</tr>
</tbody>
</table>
After recording these point estimates and variances, MPI 2.0 uses Rubin’s rule (see equations 3.9-13) to consolidate the four estimates of $\beta_2$ and make final inference. The overall point estimate of $\beta_2$ equates to 2.329 (i.e., the mean of the four point estimates) with overall variance of 0.864 (not a simple averaging result; see equation 3.10 for detail). In testing the hypothesis $H_0 : \beta_2 = 0$, the student’s $T$ statistic is 2.506 with degrees of freedom larger than 1000, and the testing p-value is less than 0.012, which is very significant. In other words, interventions involving contingency management (i.e., CMs) show stronger treatment effect than other interventions without contingency management (i.e., No-CMs). The fraction of missing information induced by intermittent missing data only is only 0.7%, which is very low. This reflects the fact that intermittent missing data alone does not account for much in the full form of the joint likelihood functions. Note that when performing MPI using the Gibbs sampler in Table 3.2, we assume that intermittent missing data are at most MAR.

**Table 3.8 Multiple estimates of $\beta_2$ for the 12-measure CO-level data**

In this chapter we discussed the ideas and algorithms of MPI and applied them to the breath data in the smoking cessation study. We also described several computational issues especially in convergence diagnostics when running the MPI Gibbs sampler. A bivariate imaginary data set was used to show how to perform EM algorithm and check convergence of the Gibbs sampler for multiple imputation.

An important point we want to re-emphasize here is the rationale for MPI. When creating the multiple imputations, we do so only for the intermittent missing data while leaving the dropouts as they originally are. Another perspective for understanding MPI, is to treat the partially imputed data as the “complete” data and ignore the dropouts as if they do not exist at all. In many cases, it is acceptable at least approximately that intermittent missingness depends only on the data that are observed. So long as the intermittent missing values are sparse and the number of repeated measures is large, the intermittent missingness usually can be well predicted by the observed data and is dependent on neither the intermittent unobserved values nor on the unobserved data due to dropout. Therefore when consolidating the multiple analyses, we found that the fraction of missing data due to intermittent missingness is often negligible. When the proportion of dropouts dominates that for intermittent missing data in a data set, like the carbon monoxide data, the assessment of dropouts and modeling strategies that can potentially handle nonignorable dropout mechanisms become more important tasks in substance abuse studies. As a computational tool for
assessing dropouts and preparing partially imputed data to be analyzed using advanced models, we believe MPI is valuable.

On the other hand, we may propose more advanced algorithms for creating partial imputations when the mechanisms of intermittent missingness are nonignorable. For example, in the case when intermittent missingness depends on dropouts, we may define the probability of dropping out as a special type of propensity score (Rosenbaum and Rubin 1983) and make partial imputations while stratifying on this score.

A final comment is that MPI offers a framework, within which more advanced models such as D-K models or P-M models can be fit to handle even nonignorable dropouts. This will be described in the next chapters.
CHAPTER 4.

SELECTION MODEL FOR NONIGNORABLE DROPOUT

In earlier chapters, we have discussed the relationships between missing data (or more specifically, missingness pattern and mechanism) and longitudinal data analysis. We have also shown how MPI can be used in assessing dropout mechanisms. In this chapter and the next two chapters, we focus on modeling strategies for incomplete longitudinal data. As mentioned in the introduction chapter, when both intermittent missing data and dropouts can be viewed as ignorable, there are plenty of modeling choices and software solutions available. For more details on longitudinal models with ignorable missing data, please refer to the introduction section and the references listed there. Note that MPI 2.0 does offer the function for fitting mixed models. Since this type of modeling assumes ignorable mechanism of missingness and dropout, it has been extensively discussed in the statistical community. We decided not to describe this modeling option in this technical report.

In this chapter, we focus on longitudinal modeling with nonignorable or informative dropouts. As seen in Chapter 3, intermittent missing data can be often viewed as ignorable (i.e., MCAR or MAR). In many substance abuse studies, e.g., smoking cessation study, intermittent missing data usually account for a small proportion of the whole data matrix. Reasons underlying intermittent missingness usually are not study related, e.g., schedule conflicts, traffic problems, and bad weather conditions. When the number of repeated measures (i.e., $J$) is large, the associated missingness mechanism for intermittent missing data can be assumed to be ignorable (Belin et al. 2000).

Dropouts, on the other hand, are of primary concern when conducting any clinical trial featuring a long period of active treatment or a treatment with intolerable adverse effects. The reasons for withdrawal are usually study-related (e.g., adverse event, death, unpleasant procedures, lack of improvement) and the mechanism underlying dropouts cannot be simply assumed to be ignorable (Verbeke and Molenberghs 2000, Yang and Shoptaw 2005).

A systematic way called sensitivity analysis will be described in this chapter within the framework of MPI where intermittent missing data are imputed using the Gibbs sampler described in section 3.2. In many cases, it is not enough to base the main data analysis only on one selected model and view the testing result as the final conclusion. Instead, a sensitivity analysis approach should be performed in order to understand how robust our final inference is to missingness assumptions.
In the following discussion, we will first introduce an important advanced incomplete longitudinal model -- selection models, and then apply it to our smoking cessation data sets to perform a sensitivity analysis. Next chapter, we will introduce another important advanced incomplete longitudinal model – pattern-mixture model.

When analyzing longitudinal data with dropouts using sensitivity analysis, the following general issues should be considered.

1. **Compliant vs. Noncompliant**: the fully compliant subgroup may be not a random sub-sample of the original total sample with different selection criteria between the treatment groups.

2. We should always compare rates, times to, and reasons for withdrawal among treatment groups.

3. Unstated reasons for dropping out of a clinical trial may be associated with the last observed response or several previous observed responses.

4. The method for handling dropout also depends on the objectives of the study, e.g., explanatory in nature during Phase I or early Phase II clinical trials, pragmatic in Phase II clinical trials.

5. It is imperative to have accurate documentation of the causes for dropout.

### 4.1 D-K Selection Model

Let us denote $t_{di}$ as the dropout time for the $i^{th}$ subject, where $2 \leq d_i \leq J + 1$ and $d_i = J + 1$ indicates a subject who has completed the study. Then, missingness indicators $R_i$ is a vector of $d_i - 1$ consecutive “0” followed by $J + 1 - d_i$ consecutive “1”. Suppressing the dependence on covariate, the selection model of Diggle and Kenward (1994) assumes: (i) $\Pr(r_{ij} = 1 | j > d_i) = 1$; (ii) for $j \leq d_i$, $\Pr(r_{ij} = 1)$ depends on $y_{ij}$ and its history $H_{ij} = (y_{i1}, ..., y_{id_i - 1})$; and (iii) the conditional distribution of $y_{ij}$ given $H_{ij}$ is $f_y(y_{ij} | H_{ij}, \theta)$. The full likelihood function for the $i^{th}$ subject is expressed as

$$L_i(\theta, \phi \mid y_{i, obs}^{obs}, R_i) \propto \prod_{j=1}^{d_i - 1} f(y_{ij} \mid H_{ij}, \theta) \prod_{j=2}^{d_i - 1} \left[ 1 - p_j(y_{ij}, H_{ij}) \right] \Pr(r_{ij} = 1 \mid H_{id_i}) \quad (4.1)$$

where $p_j(y_{ij}, H_{ij}) = \Pr(r_{ij} = 1 | y_{ij}, H_{ij}, \phi)$ indicating the probability of dropout at time $t_{ij}$. Here, we restrict that there is no dropout at the first time pint.
Dropout probability \( \Pr(r_{id_i} = 1 \mid H_{id_i}) = \int \Pr(r_{id_i} = 1 \mid y_i, H_{ij}, \theta) f_{id_i}(y \mid H_{id_i}, \theta) dy \), if \( d_i < J + 1 \); and \( \Pr(r_{id_i} = 1 \mid H_{id_i}) = 1 \), if \( d_i = J + 1 \). A natural choice for calculating \( \Pr(r_{ij} = 1 \mid y_{ij}, H_{ij}, \phi) \) is a logistic regression,

\[
\logit(\Pr(r_{ij} = 1 \mid y_{ij}, H_{ij}, \phi)) = \phi_0 + H_{ij} \phi_1 + y_{ij} \phi_2 ,
\]

where \( \phi_2 \neq 0 \) would imply an outcome-dependent nonignorable dropout process.

The full log-likelihood function of the whole data set for \((\theta, \phi)\) can be partitioned into

\[
l(\theta, \phi) = l_1(\theta) + l_2(\phi) + l_3(\phi, \theta) ,
\]

where

\[
l_1(\theta) = \sum_{i=1}^{N} \log \{ f(y_{i}^{obs}) \} \quad \text{corresponds to the observed-data log-likelihood function for } \theta ,
\]

\[
l_2(\phi) = \sum_{i=1}^{N} \sum_{j=2}^{d_i-1} \log \{ 1 - p_j(H_{ij}, y_{ij}) \} \quad \text{and}
\]

\[
l_3(\phi, \theta) = \sum_{i=N, d_i \leq J} \log \{ \Pr(r_{id_i} = 1 \mid H_{id_i}) \} \quad \text{together correspond to likelihood function for dropout process, which contains some information on } \theta . \text{ If dropouts are ignorable, then } l_1(\theta, \phi) \text{ depends only on } \phi \text{ and can therefore be absorbed into } l_2(\phi). \text{ Thus estimation of } \theta \text{ can be solely derived from } l_1(\theta).
\]

For a normal longitudinal data set, \( y_i \sim N(\chi, \Sigma(\alpha)) \) with parameters \( \theta = (\beta, \alpha) \), the conditional distribution \( f_{ij}(y \mid H_{ij}, \theta) \) is a scalar normal distribution, and the marginal distribution \( \prod_{j=1}^{d_i-1} f(y_{ij} \mid H_{ij}, \theta) = f(y_{i1} \ldots, y_{id_i-1}) = f(y_{i}^{obs}) \) is a multivariate normal distribution.

Selection models originated from the Tobit model of Heckman (1976). The theoretical translation from Tobit model to Diggle and Kenward’s selection model was addressed by Verbeke and Molenburghs (2000). Subsequently, it was extended to the non-monotone setting by Traxel, Harrington, and Lipsitz (1998). Selection models for categorical and other type of measures were also developed; see Fitzmaurice, Molenberghs, and Lipsitz (1995), Molenberghs Kenward, and Lesaffre (1997), Nordheim (1984), and Kenward and Molenburghs (1999).
4.2 Statistical Inference for D-K Models

4.2.1 Likelihood-Based Inference

To estimate the parameters \((\beta, \alpha, \varphi)\), maximize likelihood method can be used here. We should note that there are two computational difficulties in the parameters estimation by maximizing the likelihood function.

1. The integration in \(\text{Pr}(r_{id} = 1 \mid H_{id}) = \int \text{Pr}(r_{id} = 1 \mid y, H_{ij}, \varphi)f_{id} (y \mid H_{id}, \theta)dy\), cannot be performed analytically. Because \(f_{id} (y \mid H_{id}, \theta)\) is Gaussian distribution for continuous repeated measures, we can use Gauss-Hermit method to estimate integration.

2. For maximization of the overall log-likelihood (4.3), we cannot use the Newton-Raphson method, although it is much faster, because we cannot find the derivatives of \(\int \text{Pr}(r_{id} = 1 \mid y, H_{ij}, \varphi)f_{id} (y \mid H_{id}, \theta)dy\). Diggle and Kenward (1994) suggested the simplex algorithm of Nelder and Mead (1965), but there is no standard errors provided in this algorithm. MPI 2.0 applies the UNCMIN routines that are based on Appendix A of Dennis & Schnabel (1996). Same as Newton-Raphson method, starting point is very critical problem of finding the extreme value. We suggested: 1) for linear regression model parameters \((\beta, \alpha)\), we can use the parameter estimators of the mixed model omitting the missing mechanism. 2) for dropout logistic regression model parameters \(\varphi\), we can use the parameter estimators of the logistic regression model

\[
\log ii\{p_d (H_{id}, y_{id} : \varphi)\} = \phi_0 + \phi_1 y_{d-1} + \phi_2 y_{d-2} \tag{4.4}
\]

We also suggest the following general probit model for the dropout process \(P_d (H_{id}, y_{id} ; \varphi)\):

\[
p_{id} = \Phi(\phi_0 + H_{id} \varphi_1 + y_{id} \varphi_2), \tag{4.5}
\]

where \(\Phi(\cdot)\) is the cumulative density function for the standard normal distribution. Within both models, dropouts can depend on previous observed history, the present intended value.
4.2.2 Bayesian Inference

Bayesian inference based on MCMC provides an appealing alternative to the likelihood-based inferences. By sampling parameters and drawing missing values, the method of Monte Carlo using Gibbs sampler or Metropolis-Hastings algorithms offers a natural option for integration and optimization, without relying on fully-determined density functions or analytical derivatives. In the application of the Bayesian inference to the selection model, each element of the parameter vector $\psi = (\beta, \alpha, \varphi)$ is viewed as a vector of variables instead of constants, certain prior distributions $f(\psi)$ are specified, and the posterior distribution of the parameters is obtained using Bayes’ theorem,

$$P(\psi | Y, R) \propto \left[ f(y_{i}^{obs}) \prod_{j=2}^{d-1} [1 - \Pr(r_{ij} = 1 | y_{ij}, H_{ij}, \varphi)] \Pr(r_{id_i} = 1 | H_{id_i}) \right] \times f(\psi) \quad (4.6)$$

Using the method of MCMC, the posterior distribution is obtained through sampling. Inferences are then made as the summary of the statistical features of this posterior distribution, e.g., median, mean, and standard deviation.

An Augmented Gibbs Sampler

When evaluating the likelihood function model, the crucial part of computation is to calculate the actual dropout probabilities, $\Pr(r_{id_i} = 1 | H_{id_i}, \varphi)$, which requires integrating out the missing value $y_{id_i}$ over $(-\infty, +\infty)$. Only for simple cases, it is feasible to maximize the likelihood function analytically. Using an augmented Gibbs sampler, we can alleviate the computation difficulty by first imputing the missing values and then drawing parameters one by one conditionally on the observed and imputed values. More specifically, the augmented Gibbs sampler is an iterative procedure with each iteration consisting of two sub steps.

(I) Imputation-Step: where the missing values are updated by drawing from the conditional predictive distribution, that is, for $N$, draw $y_{i,d_i}^{mis}$ from

$$y_{i,d_i}^{mis} \sim f_{i,d_i}(y_i^{obs}, x_i, R, \psi) \quad (4.7)$$

For multivariate-normally distributed measures, this predictive conditional distribution is a scalar normal distribution.

(II) Estimation-Step: Where the parameters are drawn from the posterior distribution $\psi \sim P(\psi | Y^*, X, R)$ in the following order according to the decomposition of the joint distribution into full-conditional distributions,

$$\theta \sim f(\theta | (\psi, y_{1}^*, \ldots, y_{N}^*, X, R)) \quad (4.8)$$
\[ \phi \sim f(\phi \mid \psi_{\phi}, y^*_1, \ldots, y^*_N, X, R) \]  

(4.9)

where \( \theta = (\beta, \phi)^T \), \( y^*_i = (y_{i1}, \ldots, y_{i, l_{ir}})^T \), \( Y^* = (y^*_1, \ldots, y^*_N)^T \), and “\( \setminus \)” means “excluding” (e.g., \( \psi_{\setminus a} = (\beta, \phi)^T \)).

In the above algorithm, missing values are treated as another group of parameters in an approximate sense, but missing values are in fact unobserved values controlled by the parameters. They are simulated from the predictive function, instead of full-conditional distributions. In order to differentiate from standard Gibbs sampler, we call the above algorithm “augmented” Gibbs sampler. Similar ideas were adopted by Schafer (1997) in his data augmentation for creating imputations of missing values. See such an example in Chapter 3.

MPI 2.0 has implemented three versions of D-K selection models depending on the covariance structure in describing the correlation among repeated measures: AR(1) selection model, Random-Intercept selection model, and Random-Intercept and Slope selection model. For the three models, the augmented Gibbs samplers are described in Yang and Li (2005). Several practical data sets and simulated data sets verified the satisfactory performance of these MCMC algorithms. In this Technical report, however, we only give examples with likelihood inference with MLE.

### 4.3 Sensitivity Analysis Using D-K Models

In many substance abuse studies, incomplete longitudinal data analyses are usually sensitive to many modeling and missing-data assumptions. Random effects models (or more generally, mixed models) are often beyond any doubt the choice of preference to describe the measurement process should the data be complete. For incomplete data sets, they are still subject to further, sometimes untestable, modeling assumptions especially those related to missing data. Some missingness mechanisms (e.g., MCAR and DCAR) might be able to be assessed given only observed data and missingness matrices (see Chapters 1-2), but in many situations other, more complex, missingness mechanisms especially those of nonignorable type (e.g., IM or ID) are untestable.

When longitudinal models, such as D-K and P-M models, are considered to handle possibly nonignorable dropouts, several choices have to be made. For example, one has to choose between selection models and pattern-mixture models, and among various choices for dropout mechanisms.

When the D-K model, a specific form of selection model, was first proposed, it raised high expectations as seen by the published discussion to the paper (Diggle and Kenward 1994), but recently it turns out that formal tests on dropout mechanisms especially types of informative dropout, although technically
possible, should be approached with caution. The statistical community expresses growing need for methods that can investigate the sensitivity of the model-based results with respect to the model assumptions (see Nordheim 1984, Little 1994, Rubin 1994, Kenward and Molenberghs 1999, Molenberghs et al. 2001). For the same reason, in Chapter 2, we did not treat the D-K model as a tool for distinguishing between PDAR and ID (see Table 1). Instead, we discuss D-K model in this chapter and view it as a tool for sensitivity analysis.

The sensitivity study of selection models has been indicated by Glynn, Laird and Rubin (1986), Draper (1995), and Copas and Li (1997) with various definitions within different contexts. Here within the scope of this technical report, we define sensitivity analysis as one in which several statistical models along with several assumptions on missingness mechanisms are considered simultaneously for a given incomplete longitudinal data set. This definition potentially encompasses a wide range of useful approaches, although we restrict ourselves here only to D-K within the framework of MPI.

In a general sense, a whole procedure of incomplete longitudinal data analysis can be viewed as a sensitivity analysis, which is an iterative procedure that can be roughly divided into the following four steps:

(1) Preliminary data exploration and manipulation on repeated measured data and patterns of missingness (see Chapter 1).

(2) Investigation on possible missingness mechanisms for both intermittent missing data and dropouts. This step may offer valuable information to help select appropriate modeling strategies (see Chapter 2).

(3) Unless there are too many intermittent missing values and the number of repeated measures and covariates is small, perform MPI to impute these sparsely missed values (see Chapter 3).

For each partially imputed data set, fit D-K models and other advanced models (see Chapters 5 and 6) for main data analysis. Then the final conclusion is made based on these multiple analyses results using different models with different partial imputations. When different models with various assumptions consistently favor similar conclusions, we will have greater confidence in accepting these conclusions.

### 4.4 Analysis for the Carbon Monoxide Data

For the smoking cessation study, a full incomplete longitudinal data analysis was roughly divided into two parts: preliminary exploration and longitudinal modeling. In our MPI 2.0, preliminary exploration includes tasks like missing-data assessment (see Chapter 1-2) and observed data exploration (see Figures 4.1
4.4.1 Graphical Data Exploration

As plotted by MPI 2.0, Figures 4.1 and 4.2 depict the overall distribution of the observed repeated CO-levels in the smoking cessation study. The plot in Figure 4.1 is called a spaghetti-shadow plot because the light gray colored spaghetti-like profiles for all subjects offer the background (like a shadow) for the four mean curves. The mean curves jointly show the overall time-trends and relative strength of the treatment conditions, while the spaghetti-shadow background provides the variance among the observed CO-levels. Figure 4.2 is a magnified version of

Figure 4.1 Spaghetti-Shadow plot for CO-levels in the smoking cessation study (Group 1=Control, 2=RP only, 3=CM only, 4=RP+CM)

Figure 4.2 Mean CO-levels in the smoking cessation study (Group 1=Control, 2=RP only, 3=CM only, 4=RP+CM)
Figure 4.3 Spaghetti-Shadow plot for self-reported number of cigarettes in the smoking cessation study (Group 1=Control, 2=RP only, 3=CM only, 4=RP+CM)

Figure 4.4 Mean number of self-reported used cigarettes in the smoking cessation study (Group 1=Control, 2=RP only, 3=CM only, 4=RP+CM)

Figure 4.1 without the shadow background, which allows the means across treatment conditions to be more clearly seen.

Similarly, Figures 4.3 and 4.4 graphically depict the repeated self-reports of number of cigarettes in the smoking cessation study. Jointly, Figures 4.1 through 4.4 show that all four treatment conditions or interventions show similar time-trends during the 12-week study period. The most significant treatment effect occurred during the first 2 to 3 weeks and then participants maintained these levels. During most of the study, interventions involving contingency management (i.e., CM or RP+CM) correspond to lower CO-levels and smaller numbers of cigarettes used, which provides evidence in supporting their stronger treatment efficacy than other interventions (i.e., Control or RP only).

4.4.2 D-K Models for full data

From the spaghetti-shadow plot in Figure 4.1, we see that the distributions of CO-levels at each time point are asymmetric, which suggests a logarithmic
transformation. As mentioned earlier, the investigators in the smoking cessation study are most interested in the treatment effect of contingency management. Figure 4.5 depicts the spaghetti-shadow plot after such a transformation for CO-levels within group CMs and group No-CMs. It is seen clearly that during the first week the mean CO levels in both CMs and No-CMs groups drop quickly and almost linearly. Then they approximately maintain the achieved level at the end of the first week throughout the rest of the study period. Therefore, a piecewise linear mixed model can be fitted. After model selection and goodness of fit testing, we have the following model as a starting model where the dropouts are assumed to be DCAR. We call this model the D-K DCAR MODEL.

\[ Y_{ij} = \beta_0 + \beta_1 Time_1 + \beta_2 Time_2 + \beta_3 BaseCO + \beta_4 Patches + \beta_5 Group * Time_1 \] (4.10)

where \( Time_1 = \begin{cases} Time & \text{if time} \leq 3 \\ 0 & \text{if time} > 3 \end{cases}, \quad Time_2 = \begin{cases} Time & \text{if time} \leq 3 \\ Time - 3 & \text{if time} > 3 \end{cases}, \quad BaseCO \)

indicates baseline CO level, \( Patches \) represents the number of nicotine patches they received during the study (all subjects received nicotine patch), and \( Group = \begin{cases} 1 & \text{if CMs} \\ 0 & \text{if No-CMs} \end{cases} \). It is \( \beta_5 \) that is most interest, which represents the difference of the treatment effects between CMs and No-CMs. Table 4.1 shows the point estimates of \( \beta_5 \) and its standard errors across the 4 partially imputed datasets obtained using the procedure described in Chapter 3.

**Figure 4.5** Spaghetti-Shadow Plot of CO levels after log transformation within groups CMs and No-CMs
<table>
<thead>
<tr>
<th>Number of Imputation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point Estimate (( \hat{\beta}_5 ))</td>
<td>0.0946</td>
<td>0.0878</td>
<td>0.0986</td>
<td>0.0927</td>
</tr>
<tr>
<td>Std. Err. (( \sqrt{\text{var}(\hat{\beta}_5)} ))</td>
<td>0.0172</td>
<td>0.0174</td>
<td>0.0173</td>
<td>0.0173</td>
</tr>
</tbody>
</table>

**Table 4.1** Estimates of treatment-effects for each partially imputed data set using the D-K DCAR Model for the smoking cessation study

This time, let us assume that the probability of dropout at time point \( d \) (\( p_{id} ; d = 2, \ldots, 36 \)) depends only on the previously observed CO levels and possible covariates. After a similar model selection procedure, we found that the one-step preceding CO-level (i.e., \( Y_{id-1} \)) predicted the dropout probability \( p_{id} \) well. We call this model the D-K DAR Model, which is a combination of (4.20) and

\[
\log \text{it}(p_{id}) = \phi_0 + \phi_1 Y_{id-1}. \tag{4.11}
\]

Correspondingly, the estimation results on \( \beta_5 \) and \( \phi_1 \) for the four multiple partial imputed data sets are shown in Table 4.2. The values for the full log-likelihood functions are also listed in this table.

<table>
<thead>
<tr>
<th>Partial Imputations</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point Estimate (( \hat{\beta}_5 ))</td>
<td>0.0946</td>
<td>0.0877</td>
<td>0.0986</td>
<td>0.0927</td>
</tr>
<tr>
<td>Std. Err. (( \sqrt{\text{var}(\hat{\beta}_5)} ))</td>
<td>0.0258</td>
<td>0.0259</td>
<td>0.0260</td>
<td>0.0259</td>
</tr>
<tr>
<td>DAR dropout (( \hat{\phi}_0, \hat{\phi}_1 ))</td>
<td>-6.3</td>
<td>-6.3</td>
<td>-6.3</td>
<td>-6.3</td>
</tr>
<tr>
<td></td>
<td>0.802</td>
<td>0.798</td>
<td>0.795</td>
<td>0.798</td>
</tr>
<tr>
<td>Log-Likelihood</td>
<td>-17986.92</td>
<td>-18014.37</td>
<td>-18071.55</td>
<td>-18037.87</td>
</tr>
</tbody>
</table>

**Table 4.2** Estimation of treatment-effect size for each partially imputed data set assuming that dropouts are DAR.

Finally, we assume that dropouts are informative with the following dropout mechanisms

\[
\log \text{it}(p_{id}) = \phi_0 + \phi_1 Y_{id-1} + \phi_2 Y_{id}. \tag{4.12}
\]

where \( \phi_2 \) indicate the dependence of \( p_{id} \) (i.e., the probability of subject \( i \) dropping out at time \( t_{id} \)) on the missing value \( Y_{id} \). We call this model D-K ID MODEL and Table 4.3 shows the estimates for \( \beta_5 \) and dropout parameters, i.e., (\( \hat{\phi}_0, \hat{\phi}_1, \hat{\phi}_2 \)).
Comparing the log-likelihood functions between the D-K DAR MODEL and the D-K ID MODEL for each partially imputed data set, no significant differences were detected. This partially supports that the dropout probabilities depended only on the previous one-step observed CO-levels. This log-likelihood ratio test between the two types of D-K models can be also viewed as a tool to make assessment on dropout mechanisms. This can be used in Chapter 2 when testing various assumptions on dropout mechanisms. After averaging the point estimates of the dropout model (i.e., \( \hat{\phi}_0, \hat{\phi}_1, \hat{\phi}_2 \)), we have

\[
\log it(p_{id}) = -6.3 + 0.798Y_{id-1}, \tag{4.13}
\]

which suggests that the higher the one-step “previous” observed CO-level, the higher the probability a subject will dropout at the “present” time point. This is not a surprising finding since higher CO-levels correspond to worse treatment effect. When participants saw that the therapy they received did not help them, they tended to withdraw from the study.

From Tables 4.1-4.3, we see that multiple analytical results for partially imputed data sets are fairly consistent with each other. For each model, the point estimates, \( \hat{\beta}_s \)'s, are very close to each other across the four partially imputed data sets. Using the \( \hat{\beta}_s \)'s in Table 4.2 as an example, after consolidating the 4 estimates to make MPI-inferences using MPI 2.0, we have the following results which are similar to those seen in section 3.5.3.

<table>
<thead>
<tr>
<th>Number of Imputation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point Estimate (( \hat{\beta}_s ))</td>
<td>0.0946</td>
<td>0.0880</td>
<td>0.0986</td>
<td>0.0928</td>
</tr>
<tr>
<td>Std. Err. (( \sqrt{\text{var}(\hat{\beta}_s)} ))</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>ID dropout (( \hat{\phi}_0, \hat{\phi}_1, \hat{\phi}_2 ))</td>
<td>-6.757</td>
<td>-6.773</td>
<td>-6.777</td>
<td>-6.733</td>
</tr>
<tr>
<td>Std. Err.</td>
<td>0.644</td>
<td>0.691</td>
<td>0.674</td>
<td>0.642</td>
</tr>
<tr>
<td>Log-Likelihood</td>
<td>-17986.1</td>
<td>-18013.35</td>
<td>-18070.57</td>
<td>-18037.28</td>
</tr>
</tbody>
</table>

**Table 4.3.** Estimation of treatment-effect size for each partially imputed data set assuming that dropouts are non-ignorable or informative.
In testing the null hypothesis $H_0: Q = 0$

Student's T statistic = 3.539 (with 2218.64 degrees of freedom)
The testing P-value = 0.0004

The fraction of missing information for this parameter: 0.0376

In this post-hoc longitudinal analysis within the framework of MPI, the treatment effect of contingency management (CMs) was found to be significantly better than interventions without it (i.e., No-CMs) for smoking cessation purpose in this methadone maintained study population ($T_{2218.64} = 3.539, p=0.0004$ in the D-K DAR MODEL).

### 4.4.3 D-K Models for partial data

After looking at the graph of the mean curves of the carbon monoxide levels, we found there was a big change from week 1 to week 2. Here, we mainly investigated the effect of dropout. There were only 2 observations whose dropout time was in the 2nd week. We deleted that 2 observations and use the data without the observation of first week.

![Plot 4.6](image)

**Figure 4.6** Mean CO levels of smoking data. (CM= red; Non-CM = green)

After looking the mean profile plot (figure 4.6), a linear mixed model can be fitted. After model selection and goodness of fit testing, we have the following model as a starting model where the dropouts are assumed to be ID. First, fit the mixed model, after model selection, we have model like:

$$Y_{ij} = \beta_0 + \beta_1 \ast CM + \beta_2 \ast PATCHES + \beta_3 \ast BASELINE + \beta_4 \ast Time + \epsilon_{ij} \quad (4.14)$$

with random intercept and slope, which are absorbed into the residual errors.
For the dropout logistic regression, we use (4.5), but the estimation of $\phi_1$ and $\phi_2$ (Table 4.4) are not significant at significant level 0.05.

<table>
<thead>
<tr>
<th>Number of Imputation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Point Estimate ($\hat{\phi}_1$)</strong></td>
<td>0.036</td>
<td>0.021</td>
<td>0.030</td>
<td>0.032</td>
</tr>
<tr>
<td>Std. Err ($\sqrt{\text{var}(\hat{\phi}_1)}$)</td>
<td>0.025</td>
<td>0.028</td>
<td>0.026</td>
<td>0.026</td>
</tr>
<tr>
<td><strong>Point Estimate ($\hat{\phi}_2$)</strong></td>
<td>0.024</td>
<td>0.036</td>
<td>0.029</td>
<td>0.027</td>
</tr>
<tr>
<td>Std. Err ($\sqrt{\text{var}(\hat{\phi}_2)}$)</td>
<td>0.023</td>
<td>0.024</td>
<td>0.023</td>
<td>0.023</td>
</tr>
</tbody>
</table>

**Table 4.4.** Estimation of parameters of dropout logistic regression for each partially imputed data set assuming that dropouts are non-ignorable.

Then we set the dropout logistic model as MAR with the formula:

$$\logit(p_{id}) = \phi_0 + \phi_1 Y_{id-1}$$

(4.15)

The estimation of $\phi_1$ (table 4.5) is significant bigger than zero.

After averaging the point estimates of the dropout model (i.e., $\hat{\phi}_0, \hat{\phi}_1$), we have

$$\logit(p_{id}) = -3.79 + 0.052 Y_{id-1},$$

(4.16)

which suggests that the higher the one-step “previous” observed CO-level, the higher the probability a subject will dropout at the “present” time point. The result is consistent with result of Section 4.4.2.

<table>
<thead>
<tr>
<th>Number of Imputation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Point Estimate ($\hat{\phi}_0$)</strong></td>
<td>-3.7845</td>
<td>-3.7901</td>
<td>-3.7875</td>
<td>-3.7856</td>
</tr>
<tr>
<td>Std. Er. ($\sqrt{\text{var}(\hat{\phi}_0)}$)</td>
<td>0.2120</td>
<td>0.2117</td>
<td>0.2114</td>
<td>0.2120</td>
</tr>
<tr>
<td><strong>Point Estimate ($\hat{\phi}_1$)</strong></td>
<td>0.0513</td>
<td>0.0522</td>
<td>0.0518</td>
<td>0.0516</td>
</tr>
<tr>
<td>Std. Err. ($\sqrt{\text{var}(\hat{\phi}_1)}$)</td>
<td>0.0132</td>
<td>0.0132</td>
<td>0.0132</td>
<td>0.0132</td>
</tr>
</tbody>
</table>

**Table 4.5.** Estimation of parameters of dropout logistic regression for each partially imputed data set assuming that dropouts are PDAR.
4.5 Analyzing the Milk Protein Data

The data from the milk protein trial are taken from Verbyla and Cullis (1990). They consist of assayed protein content of milk samples taken weekly from each of 79 cows. The cows were randomly allocated to one of three diets: barley, mixed barley-lupins and lupins, with 25, 27 and 27 animals in the three groups. Measurements were taken for up to 19 weeks, but there were 38 drop-outs from week 15 onwards, corresponding to cows who stopped producing milk before the end of the experiment. The primary objective of the experiment was to describe the effects of diet on the mean response profile over time.

4.5.1 Previous Study for Milk Protein Data

As plotted by MPI 2.0, Figure 4.7 depicts the overall distribution of the observed response in the milk protein study. The mean curves jointly show the overall time-trends and relative strength of the treatment conditions, while the spaghetti-shadow background provides the variance among the observed responses.

From Figure 4.7, the observed mean response profiles are almost parallel except the last five weeks. We should note that from week 15 onwards the means are calculated only from cows still producing milk.

![Figure 4.7 Spaghetti-Shadow plot for response profiles for the milk protein data: (red= barley, green= mixed barley-lupins, blue= lupins).](image)
Previous analyses without consideration of the dropout mechanism for these data are reported by Diggle (1990) and Verbyla and Cullis (1990).

Diggle analyzed the entire data set under the implicit assumption of PDAR. Verbyla and Cullis ignored the data from the first three weeks, use 16 parameters to describe the mean response $\mu_{t,j}, j=1,\ldots,16$, for diet 1 in each of the last weeks and describe the contrasts $\mu_{2,j} - \mu_{1,j}$ and $\mu_{3,j} - \mu_{1,j}$. They all conclude that there are significant differences between the three responses profile.

Diggle and Kenward (1994) analyze this data use the D-K model and the dropout with the logistic expression

$$
\log \text{it}\{p_d(H_d, y_d : \phi)\} = \phi_{0,d-1} + \phi_{1}y_{d-1} + \phi_{2}y_{d} \quad d = 15,\ldots,19
$$

(4.17)

because the observed mean response profiles tend to rise towards the end of the experiment. Diggle and Kenward’s result (considering about the dropout mechanism) is different to the result of Diggle (1990; without considering the dropout mechanism). And they also concluded that the dropout out missing mechanism was ID.

### 4.5.2 Re-analysis Using a New Selection Model

Now, we re-analyze the study using the data after ignoring the first 6 weeks. The new data have up to 13 measurements. We assume that the probability of dropout at time point $d$ depends only on the present and previously measurements.

Our model combines a linear model for repeated measures:

$$
Y_g = \beta_0 + \beta_1 * Time + \beta_2 * cond1 + \beta_3 * cond2 + \\
\beta_4 * Time * cond1 + \beta_5 * Time * Cond2 + \varepsilon_g
$$

(4.18)

with CS (compound symmetric) covariance structure, and a logistic dropout model for dropout mechanism:

$$
\log \text{it}\{p_d(H_d, y_d : \phi)\} = \phi_{0} + \phi_{1}y_{d-1} + \phi_{2}y_{d},
$$

(4.19)

where $Cond1 = \begin{cases} 1 & \text{if mixed barley-lupins} \\ 0 & \text{others} \end{cases}$, $cond2 = \begin{cases} 1 & \text{if lupins} \\ 0 & \text{others} \end{cases}$, and $d$ is the dropout time. To test the null hypothesis—“there is no significant difference between the three responses profile,” it is $\beta_4$ and $\beta_5$ that are of most interest,
which represent the differences of the treatment effects between the three diets. Table 4.6 shows the point estimates of $\beta_4$, $\beta_5$ and their standard errors across the 4 partially imputed datasets obtained using the procedure described in Chapter 3.

<table>
<thead>
<tr>
<th>Number of Imputation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point Estimate ($\hat{\beta}_4$)</td>
<td>0.00110</td>
<td>0.00084</td>
<td>0.00022</td>
<td>0.00092</td>
</tr>
<tr>
<td>Std. Err. ($\sqrt{\text{var}(\hat{\beta}_4)}$)</td>
<td>0.00597</td>
<td>0.00593</td>
<td>0.00595</td>
<td>0.00593</td>
</tr>
<tr>
<td>Point Estimate ($\hat{\beta}_5$)</td>
<td>-0.0058</td>
<td>-0.0056</td>
<td>-0.0060</td>
<td>-0.0054</td>
</tr>
<tr>
<td>Std. Err. ($\sqrt{\text{var}(\hat{\beta}_5)}$)</td>
<td>0.00601</td>
<td>0.00597</td>
<td>0.00599</td>
<td>0.00597</td>
</tr>
<tr>
<td>Log-Likelihood</td>
<td>-173.2</td>
<td>-166.5</td>
<td>-170.2</td>
<td>-166.6</td>
</tr>
</tbody>
</table>

Table 4.6: Estimates of treatment-effects for each partially imputed data set using the D-K model (4.18-19) for the milk protein study.

It is seen that none of the estimates is significant. Then we modified the model of the mean structure as

$$E(Y_i) = \beta_0 + \beta_1 \times \text{Time} + \beta_2 \times \text{cond1} + \beta_3 \times \text{cond2}. \quad (4.20)$$

The same logistic dropout model as (4.19) was used. The estimated parameters are listed in Table 4.7.

<table>
<thead>
<tr>
<th>Number of Imputation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point Estimate ($\hat{\beta}_0$)</td>
<td>3.48</td>
<td>3.48</td>
<td>3.47</td>
<td>3.47</td>
</tr>
<tr>
<td>Point Estimate ($\hat{\beta}_1$)</td>
<td>0.0082</td>
<td>0.0084</td>
<td>0.0084</td>
<td>0.0085</td>
</tr>
<tr>
<td>Std. Err. ($\sqrt{\text{var}(\hat{\beta}_1)}$)</td>
<td>0.0025</td>
<td>0.0024</td>
<td>0.0025</td>
<td>0.0025</td>
</tr>
<tr>
<td>Point Estimate ($\hat{\beta}_2$)</td>
<td>-0.106</td>
<td>-0.108</td>
<td>-0.102</td>
<td>-0.105</td>
</tr>
<tr>
<td>Std. Err. ($\sqrt{\text{var}(\hat{\beta}_2)}$)</td>
<td>0.054</td>
<td>0.053</td>
<td>0.053</td>
<td>0.053</td>
</tr>
<tr>
<td>Point Estimate ($\hat{\beta}_3$)</td>
<td>-0.234</td>
<td>-0.237</td>
<td>-0.230</td>
<td>-0.235</td>
</tr>
<tr>
<td>Std. Err. ($\sqrt{\text{var}(\hat{\beta}_3)}$)</td>
<td>0.054</td>
<td>0.053</td>
<td>0.053</td>
<td>0.053</td>
</tr>
<tr>
<td>Point Estimate ($\hat{\phi}_0$)</td>
<td>2.4</td>
<td>1.8</td>
<td>2.0</td>
<td>1.7</td>
</tr>
<tr>
<td>Point Estimate ($\hat{\phi}_1$)</td>
<td>-11.3</td>
<td>-11.4</td>
<td>-11.5</td>
<td>-11.1</td>
</tr>
<tr>
<td>Std. Err. ($\sqrt{\text{var}(\hat{\phi}_1)}$)</td>
<td>1.8</td>
<td>1.8</td>
<td>1.9</td>
<td>1.9</td>
</tr>
<tr>
<td>Point Estimate ($\hat{\phi}_2$)</td>
<td>8.8</td>
<td>9.0</td>
<td>9.0</td>
<td>9.1</td>
</tr>
<tr>
<td>Std. Err. ($\sqrt{\text{var}(\hat{\phi}_2)}$)</td>
<td>1.5</td>
<td>1.6</td>
<td>1.6</td>
<td>1.6</td>
</tr>
<tr>
<td>Log-Likelihood</td>
<td>-174.0</td>
<td>-167.1</td>
<td>-170.9</td>
<td>-167.3</td>
</tr>
</tbody>
</table>

Table 4.7: The result for the new model (4.19 & 4.20) for the milk protein study.
When comparing the above two D-K models, likelihood ratio test was used. In testing the hypothesis of \( \beta_4 = \beta_5 = 0 \), likelihood-ratio chi-square statistics for all the four imputed data are all less than 1.5 with 2 degree of freedom, which are not significant. Therefore, we concluded that the interaction between time and treatment condition is not important. That is, there is no significant difference between the three response profiles on slope.

The linear regression part of our model was:

\[
Y_{ij} = 3.48 + 0.0084 \cdot \text{Time} - 0.105 \cdot \text{cond1} - 0.234 \cdot \text{cond2}
\]

which suggested that the response was increasing with time and the mean response of three diets were different. But the difference did not change with the time.

Then we checked the estimate of \( \phi_1 \) and \( \phi_2 \), both were very significant, so we concluded that the drop-out process was ID. Interestingly, and at first sight paradoxically, the maximum likelihood estimates of \( \phi_1 \) and \( \phi_2 \) had opposite signs, \( \phi_1 = -11.3 \) and \( \phi_2 = 9 \). By adopting a re-parameterization of the dropout parameters, we obtained an easier interpretation of the result.

By writing equation 4.19 into

\[
\log \{ P_d (H_d, y_d \mid \phi) \} = \phi_0 + \theta_1 (y_d + y_{d-1}) + \theta_2 (y_d - y_{d-1})
\]

with a set of new parameters, \( \theta_1 = (\phi_1 + \phi_2) / 2 \) and \( \theta_2 = (\phi_2 - \phi_1) / 2 \), we could interpret \( \theta_1 \) and \( \theta_2 \) easily. They represent dependence on level and increment in the response variable, and these quantities were likely to be much less correlated than \( y_d \) and \( y_{d-1} \). The maximum likelihood estimates of the \( \theta_1 \) and \( \theta_2 \) are \( \hat{\theta}_1 = -1.2 \) and \( \hat{\theta}_2 = 10.3 \), suggesting that the probability of drop-out increases when either the prevailing level of protein content is low or when the increment between the last and current protein contents is high.

Finally, we noted that our detailed consideration of the dropout process for these data has not led to any material change in the conclusion about the complete measurement process \( Y^* \). From a theoretical point of view, this is not surprising because the drop-outs were confined to the last five weeks of the experiment, and the dropout-free phase of the experiment contains most of the information about the \( Y^* \) process. From a practical point of view, we think that it has been useful to expose the distinction between the mean response profiles for \( Y^* \) and for \( Y \) conditional on non-dropout.
4.6 A Simulation Study

Diggle and Kenward (1994) did some simple simulation studies to make a comparison with the simpler PDAR analysis for data generated under an ID process and compare the behavior of OSL (ordinary least squares) estimators. They concluded: “when data are generated under an ID process, the ML estimators from the ID model do not suffer from the bias that is present in the OSL and RD (i.e., PDAR) ML estimators.” In their simulation, they used a simple linear model, which was too simple to be practically realistic.

We did a simulation study to check the performance of this model with more complicated linear models and more structures of the variance covariance matrix. We specified three aspects of the models to be simulated: the linear mean structure, the covariance structure, and the dropout model.

First, we define a linear model as

\[ Y_{ij} = \beta_0 + \beta_1 \cdot \text{cond}_{ij} + \beta_2 \cdot \text{time}_{ij} + \beta_3 \cdot X_{ij} + \beta_4 \cdot X_{2i} + \beta_5 \cdot \text{time}_{ij} \cdot \text{cond}_{ij} + \epsilon_{ij} \]  

(4.23)

where \( \beta_0 = 2, \beta_1 = 0, \beta_2 = 1, \beta_3 = 0.5, \beta_4 = 0, \) and \( \beta_5 = -0.5. \) Two conditions are used: \( \text{cond}_{ij} = 0 \) or 1, respectively representing a control or treatment condition.

Second, the covariance structure is specified as a first-order auto-regression (AR(1)) structure or a compound symmetric (CS) structure.

Third, we allowed the dropout model to be DCAR, PDAR, or ID, where the probability of dropout is calculated from the logistic regression model:

\[ \text{Logit} \ P_{id} = \phi_0 + \phi_1 y_{i,d-1} + \phi_2 y_{id} \]  

(4.24)

and \( \phi_1 = \phi_2 = 0 \) corresponds to DCAR, \( \phi_2 = 0 \) corresponds to PDAR, ID corresponds to the case of no restriction on \( \phi_1 \) and \( \phi_2. \) Value of \( \phi_0, \phi_1 \) and \( \phi_2 \) were chosen to produce proportions of missing values equal to approximately 30%.

Eventually, we have the following factors in this design of the experiment: dropout mechanism (DAR, DCAR or ID), lag-one correlation between subject (\( \rho = 0.5 \) or 0.8), subject number in each conditions (50 or 100), and number of repeated measure for each subject (5 or 10). For each combination of the above design factors, 100 sets of complete data were generated from a multivariate-normal random number generator and then dropouts (missing values due to dropout) are created by deleting data points using the chosen dropout mechanism.
For each set of simulated data, the D-K model with linear model (4.23) and the dropout logistic model (4.25) are used to estimate the parameters.

Since the comparison of trends over time frequently arises in the analysis of longitudinal data, we first wanted to check whether $\beta_5 = -0.5$, which is the difference of the trends over time between two conditions.

<table>
<thead>
<tr>
<th>Total number of observation</th>
<th>Maximum repeated measure</th>
<th>$\rho$ Lag1 correlation</th>
<th>Drop-out mechanism</th>
<th>Estimator using different missing mechanism</th>
<th>Average % missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>5</td>
<td>.5</td>
<td>DCAR</td>
<td>ID 6 11 12</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PDAR</td>
<td>ID 2 7 7</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.8</td>
<td>DCAR</td>
<td>ID 2 7 7</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PDAR</td>
<td>ID 2 7 7</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.8</td>
<td>DCAR</td>
<td>ID 2 7 7</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PDAR</td>
<td>ID 2 7 7</td>
<td>30</td>
</tr>
<tr>
<td>200</td>
<td>5</td>
<td>.5</td>
<td>DCAR</td>
<td>ID 2 17 17</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PDAR</td>
<td>ID 2 17 17</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.8</td>
<td>DCAR</td>
<td>ID 2 17 17</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PDAR</td>
<td>ID 2 17 17</td>
<td>31</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>.5</td>
<td>DCAR</td>
<td>ID 2 8 8</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PDAR</td>
<td>ID 2 8 8</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>.8</td>
<td></td>
<td>DCAR</td>
<td>ID 2 8 8</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PDAR</td>
<td>ID 2 8 8</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 4.8 Percentage bias in the estimate of slope difference from the simulation data sets.

For each set of simulated data, the DCAR, PDAR and ID selection models were fitted by using maximum likelihood estimation (as described in section 4.2.1). The relative bias in testing $H_0: \beta_5 = -0.5$ with significant level $\alpha = .05$ is presented in Table 4.8 as a percentage for the case of AR(1). Bias is defined as rejecting the null hypotheses, $H_0: \beta_5 = -0.5$. We know, if the model is correct, the value in the table should be around 5%.
From the table, if the data is generated by PDAR (or DCAR) drop-out mechanism, there is no much different when we analyze the data using three different drop-out mechanism. But when the data is generated by informative drop-out mechanism, only the estimators getting from the ID drop-out mechanism have the bias has around 5% and the estimators getting from the MAR and MCAR drop-out mechanism have the bias bigger than 5% and up to 17%.

Here we only showed the results for the AR(1) covariance structure. Results for the case of CS looks similar, thus skipped for display.

**SUMMARY FOR CHAPTER 4**

In this chapter, we proposed a general procedure for incomplete longitudinal data analysis adopting the idea of sensitivity analysis within the framework of MPI. One special selection model, D-K model, was introduced. Then a sensitive study and simulation study was addressed. At the end, D-K models was applied to the smoking cessation data sets and milk protein data sets, which jointly supported an acceptable performance of D-K model with likelihood-based inferences.

A difficulty in fitting a D-K model using UNCMIN routines is its slow convergence speed. For example, in fitting our D-K ID MODEL to the carbon monoxide level data, it took as long as ten minutes for the algorithm to converge. Another limitation for this algorithm is that this algorithm is very sensitive to the starting point, so how to find a good starting point is a big problem. As an alternative solution, Yang and Li (2005) proposed the MCMC-based fitting algorithms for making Bayesian inferences. The augmented Gibbs sampler, simulation results, and practical application are summarized in the manuscript and the method has been implemented into MPI 2.0.

As commented by Molenberghs, Kenward & Lesaffre (1997), “conclusions are conditional on the appropriateness of the assumed model, which in a fundamental sense is not testable”, we should be cautious when using D-K model in practice. At least other modeling strategies (introduced in the following chapters) should be applied to investigate how sensitivity the fixed parameter estimates are to the assumptions on missing values.
CHAPTER 5.

PATTERN-MIXTURE MODEL FOR NONIGNORABLE DROPOUT

The high sensitivity of selection modeling has led to a growing interest in pattern-mixture modeling (Glynn, Laird, and Rubin 1986; Thijs et al., 2002). After initial introduction (Litte 1993, 1994), they are receiving more attention lately for continuous repeated measures (e.g., Little, 1995; Ekholm and Skinner, 1998; Hogan and Laird, 1996; Molenberghs et al., 1998; and Michiels et al., 1999) and for categorical measures (e.g., Molenberghs, Michiels, and Lipsitz, 1999; Birmingham and Fitzmaurice, 2002; Birmingham, Rotnitzky, and Fitzmaurice, 2003). In this chapter, two strategies for applying a pattern-mixture model along with practical applications are described.

5.1 P-M Models

Litttle (1993, 1994, 1995) has developed a general class of models called “pattern-mixture models” to deal with missing data where subjects are divided into groups classified by their missingness patterns. The idea of pattern-mixture was seen in Glynn, Laird, and Rubin (1986) and Marini, Olsen, and Rubin (1980). It is also used in other specific statistical models such as linear regression (Cohen and Cohen 1983), structure equation models (Allison 1987, Muthen, Kaplan, and Hollis 1987), and random effects models (Hogan and Laird 1997). In order to contrast these with the D-K model described in section 4.1, we call pattern-mixture models P-M models.

As mentioned earlier in Section B4 of the Introduction, the full-likelihood function should be used in modeling the incomplete longitudinal data,

\[
L(\theta, \phi \mid y_{i,obs}, r_i) \propto \prod_{i=1}^{N} \int f(y_{i, r_i} \mid \theta, \phi) dy_{i, mis}.
\]  

(5.1)

In the previous chapter, based on the outcome-dependent missingness, the selection models were addressed. Here in this chapter, the case of P-M models is discussed. A P-M model is a pattern-dependent model assuming that distribution of repeated measures varies with the missingness patterns and the joint distribution is factored as

\[
f(y_{i, r_i} \mid X_i, \theta, \phi) = f(y_{i, r_i, X_i} \mid \theta, \phi) f(r_i \mid X_i, \phi).
\]  

(5.2)
Assuming that there are \( P \) patterns of missingness in a data set, the marginal model of \( y_i \) is a mixture model,

\[
f(y_i) = \sum_{p=1}^{P} f(y_i | r_i = p, X_i, \theta^{(p)}) \pi_p
\]

where \( \theta^{(p)} \) represents the parameters of \( f(y_i) \) in the \( p^{th} \) pattern, \( \pi_p = \frac{\text{Pr}(r_i = p | X_i, \phi)}{\sum_{p=1}^{P} \frac{\text{Pr}(r_i = p | X_i, \phi)}} \), and \( r_i \) is replaced by a scalar \( r_i \) to numerate the \( P \) patterns (\( p = 1, \ldots, P \)).

It should be emphasized here that \( \theta^{(1)}, \ldots, \theta^{(P)} \) may be different both in dimensionality and in value. Accordingly, there are two strategies in the parameter estimation and model fitting. Little (1993, 1994) advocated the use of identifying restrictions which works well in relatively simple settings. This idea was further extended by Molenberghs, Michiels, and Diggle (1998) and Thijs et al. (2002) so that more general practical data sets can be analyzed. This idea corresponds to the case of same-dimension parameters (i.e., \( \theta^{(1)}, \ldots, \theta^{(P)} \)) with changing values.

The second strategy is to adopt simplified models, which corresponds to the case of varying-dimensional parameters. In this strategy, the number of total parameters is decreased, thus the difficulty of un-identification is removed, but the simplified models associate with stronger assumptions.

### 5.2 P-M Models with Identified Parameters

The articles by Little (1995) and Demitras and Schafer (2002) provide statistically rigorous treatment of random-effects P-M models for longitudinal data with dropouts that are not ignorable. Adopting similar notation used earlier in this technical report, let us first give a typical random effects model for clinical trial data:

\[
y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 (x_i \times t_{ij}) + \gamma_{ui} + \gamma_{ti} + e_{ij}.
\]

This is a random intercept \( (\gamma_{ui}) \) and slope \( (\gamma_{ti}) \) model where \( x_i \) represents the covariate of interest, e.g., treatment condition, and \( t_{ij} \) \( (i = 1, \ldots, N, j = 1, \ldots, J) \) represents the measured time for \( y_{ij} \).

Within the framework of MPI where intermittent missing data have been imputed, we apply P-M models in three steps:
(1) Classify subjects in a data set into strata depending on their patterns of missingness. These strata then are used to examine the effect of missing-data pattern on the outcomes of interest.

(2) Within each stratum, a random effects model is fitted to estimate treatment effects (e.g., $\beta_3$) and time trend (e.g., $\beta_1$).

(3) Combine the within-stratum analyses together to make final inferences.

A even simplified P-M procedure merges the above second and third steps into one step by adding variables indicating missingness patterns to the right side of model (5.4) as predictors. For example, if we roughly divide the whole data set into two parts: completers (who did not withdraw during the study period) vs. dropouts (who did withdraw), we can use a binary variable called DROP to represent this missingness patterns (=1 if dropout; =0 otherwise). By adding DROP into (5.4), we have the following form of P-M model:

$$
Y_j = \beta_0 + \beta_1 t_j + \beta_2 X_j + \beta_3 (X_j \times t_j) + \beta_4 \text{DROP}_j + \beta_5 (\text{DROP}_j \times t_j) + \beta_6 (X_j \times \text{DROP}_j) + \beta_7 (X_j \times \text{DROP}_j \times t_j) + \gamma_0 + \gamma_1 t_j + \epsilon_j
$$

(5.5)

Within P-M models, dropout mechanisms do not have obvious definitions as we did for selection models (see Table 1 in the introductory section B3). When specific modeling of the dropout mechanism is required, please refer to the rigorous theory given by Little (1995) and a systematic simulation study performed by Demitras and Schafer (2002). On the application side, Hedeker and Gibbons (1997) describe in detail how to apply these models with identified parameters to the psychological longitudinal data.

Later in this chapter, the above strategy with simplified models is applied to fit a P-M model to the smoking cessation carbon monoxide data set.

5.3 P-M Models with Identification of Restrictions

A pattern-mixture model for dropouts factorizes the joint distribution $f(y_i, d_i | \theta, \phi)$ into the product of the marginal distribution $f(d_i | X_i, \phi)$ and the conditional distribution $f(y_i | X_i, \theta^{(d,i)})$, where $d_i = 2, \ldots, J+1$ indicating the dropout time. The difficulty in pattern-mixture modeling of premature dropouts regards parameter identification. For any subject with $d_i < J+1$, the sub-vector of $\theta^{(d,i)}$ describing $y_i^{mis}$ is generally unidentified, unless certain restrictions are applied. Thijs et al. (2002) proposed a framework for identifying restrictions. By suppressing the subscript “$i$” and using $j = 1, \ldots, J$ (i.e., $j = d - 1$) to indicate all
the $J$ possible dropout patterns, the complete data density for pattern $j$ is given by

$$f_j(y) = f_j(y^{\text{obs}})f_j(y^{\text{mis}} | y^{\text{obs}}),$$  \hspace{1cm} (5.6)

where $y^{\text{obs}} = (y_1, \ldots, y_J)^T$, and $f_j(y^{\text{mis}} | y^{\text{obs}})$ is the density for the conditional distribution, which cannot be identified within this pattern. By borrowing information from observed data in other patterns where $y_s \in y^{\text{mis}}$ is observed, it is possible that $f_j(y^{\text{mis}} | y^{\text{obs}})$ be identified. By introducing some proper weights (i.e., $\sum_{j=1}^J \omega_j = 1$), we can identify $f_j(y_s | y_1, \ldots, y_{s-1})$ by

$$f_j(y_s | y_1, \ldots, y_{s-1}) = \sum_{j=1}^J \omega_j f_j(y_s | y_1, \ldots, y_{s-1}), \hspace{0.5cm} s = j + 1, \ldots, J. \hspace{1cm} (5.7)$$

Using this restriction method, the full density can be expressed as

$$f_j(y) = f_j(y^{\text{obs}}) \prod_{s=0}^{J-1} \left[ \sum_{j=1}^J \omega_{T-s,j} f_j(y_{T-s} | y_1, \ldots, y_{T-s-1}) \right]. \hspace{1cm} (5.8)$$

Depending on the specification of $\omega_{st}$’s, various schemes of identification can be implemented. For example, if all the weights are set to positive values, this will correspond to the identification using “available case missing values” (ACMV; Molenburghs et al., 1998). It is of special interest to us because it is the natural counterpart of MAR in the outcome-dependence missingness framework.

The restriction called “complete-cases missing variable” (CCMV; Little, 1993) identifies $f_j(y_s | y_1, \ldots, y_{s-1})$ by borrowing information only from the completers, i.e.,

$$f_j(y_s | y_1, \ldots, y_{s-1}) = f_j(y_s | y_1, \ldots, y_{s-1}), \hspace{0.5cm} s = j + 1, \ldots, J, \hspace{1cm} (5.9)$$

which is a special case of (5.7) with $\omega_{st} = 1$ and $\omega_{st} = \omega_{s,s+1} = \ldots = \omega_{s,T-1} = 0$.

Another special case of identification is via “neighboring case missing values” (NCMV), which borrows information from neighbors with observed values on $y_s$, i.e.,

$$f_j(y_s | y_1, \ldots, y_{s-1}) = f_j(y_s | y_1, \ldots, y_{s-1}), \hspace{0.5cm} s = j + 1, \ldots, J, \hspace{1cm} (5.10)$$

which corresponds to $\omega_{ss} = 1$ and $\omega_{s,s+1} = \omega_{s,s+2} = \ldots = \omega_{s,T} = 0$. 

97
The above identification approach based on (5.7) is convenient for sensitivity analysis. By varying the specification of weights (i.e., $\omega_u$), it is equivalent to specify different assumptions on the dropout mechanism. When the number of dropout patterns is large, it would be tedious that we apply the above identification strategy directly using a likelihood-based method to estimate the parameters for each pattern and then combine them. As mentioned earlier, the method of imputation would be a convenient tool.

5.4 Identification of Restrictions within the framework of MPI

For incomplete longitudinal data sets, the method of multiple imputation (Rubin, 1987) is especially useful. Accurate predictions of missing values because repeated measures are often highly correlated to each other. When imputing, all above three modeling options can be used. In longitudinal data sets, missingness patterns and mechanisms for intermittent missing values and dropouts are apt to be distinct, thus requiring different treatment. Our empirical experiences suggest that in certain clinical trials intermittent missing values are ignorable due to factors that are non-related to the theme of the study, while dropouts should not be simply ignored. In Yang and Shoptaw (2005), a partial version of multiple imputation, MPI, was first proposed, within which only intermittent missing values are imputed. As seen in the application of pattern-mixture model, imputation methods can be further employed to implement various schemes of identification of restriction for managing dropouts. This leads to a further extension of multiple imputation, which we term 2-stage MPI here. Depending on the assumptions of the mechanism of intermittent missingness and dropout, there exist many specific forms of MPI and 2-stage MPI. MCMC algorithms for creating imputations for ignorable intermittent missing values were described in Chapter 3. Here, in this chapter, the idea of MPI is further generalized so that, along with the 2-stage MPI, it provides a framework for sensitivity analysis.

5.4.1 MPI and 2-Stage MPI

We further partitioned $y_{i \text{mis}}$ into $(y_{i}^{IM}, y_{i}^{DM})$ to denote intermittent missing values and dropouts. For MPI, we draw $m > 1$ independent values $y_{i}^{IM(1)}, y_{i}^{IM(2)}, \ldots, y_{i}^{IM(m)}$ using the posterior predictive distribution $p(y_{i}^{im} \mid y_{i}^{obs}, r_{j})$. Within the 2-stage MPI, for each of the partial imputation for intermittent missing values, $n$ conditionally independent values $y_{i}^{DM(j,1)}, y_{i}^{DM(j,2)}, \ldots, y_{i}^{DM(j,n)}$ are additionally drawn from the predictive distribution $p(y_{i}^{DM(j)} \mid y_{i}^{obs}, y_{i}^{IM(j)})$, $j = 1, \ldots, m$. As mentioned earlier, the 2-stage MPI provides a natural framework for fitting
pattern-mixture models by identifying restrictions with information borrowed from completers, neighboring cases, or available cases. If we use selection models or REMTMs (see Chapter 6), imputations for dropouts can be similarly conducted by applying appropriate MCMC algorithms.

5.4.2 Consolidating Results from Post-Imputation Analyses

A main concern for multiple imputation is how to combine the multiple point estimators to make an overall inferential statement. A set of rules for combination was originally developed by Rubin and Schenker (1986), which can be used directly for MPI. (For the one stage MPI, the set of rules was described in Chapter 3.)

In Shen (2000), the idea was extended for the case of 2-step multiple imputation, which can be viewed as a general setting for 2-stage MPI. More specifically, \( m \times n \) complete data sets are obtained eventually in the 2-stage MPI, \( y_{ij}^{(j,k)} = (y_{ij}^{obs}, y_{ij}^{DM(j)}, y_{ij}^{DM(j,k)}) \): \( j = 1, \ldots, m \), \( k = 1, \ldots, n \). A noticeable problem with these complete data sets is that they are not independent from each other, because each block or nest \( y_{ij}^{DM(j,1)}, y_{ij}^{DM(j,2)}, \ldots, y_{ij}^{DM(j,n)} \) contains identical values for \( y_{ij}^{DM(j)} \). By denoting \( \hat{Q}^{(j,k)} \) and \( \sqrt{U}^{(j,k)} \) as the point and variance estimates for \( Q \) from the \( j, k \)th completed data set, the overall point estimate for \( Q \) is still the simply grand average, i.e.,

\[
\overline{Q} = \frac{1}{mn} \sum_{j=1}^{m} \sum_{k=1}^{n} \hat{Q}^{(j,k)}.
\] (5.11)

The associated variance for \( \overline{Q} \) involves three components, i.e.,

\[
T = \overline{U} + (1 - \frac{1}{n})W + (1 + \frac{1}{m})B
\] (5.12)

where \( \overline{U} = \frac{1}{mn} \sum_{j=1}^{m} \sum_{k=1}^{n} U^{(j,k)} \) estimates the complete-data variance,

\[
B = \frac{1}{m-1} \sum_{j=1}^{m} (\overline{Q}^{(j,\cdot)} - \overline{Q})
\] indicates the between-nest variance,

\[
W = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{n-1} \sum_{k=1}^{n} (\hat{Q}^{(j,k)} - \overline{Q}^{(j,\cdot)})^2
\] represents the within-nest variance, and

\[
\overline{Q}^{(j,\cdot)} = \frac{1}{n} \sum_{k=1}^{n} \hat{Q}^{(j,k)}.
\] Inferences about \( Q \) are based on the Student’s t-distribution

\[
\frac{(Q - \overline{Q})}{\sqrt{T}} \sim t_v \quad \text{with d.f.} \quad v = \frac{1}{m(n-1)} \left[ \frac{(1-1/n)W}{T} \right]^2 + \frac{1}{m-1} \left[ \frac{(1+1/m)W}{T} \right]^2.
\] Other formulas such as rates of missing information and relative efficiency are seen in

5.4.3 Creating Imputations in the MPI and the 2-stage MPI:

In Chapter 6, the hybrid Gibbs sampler is proposed to make Bayesian inferences using a Shared-Parameter model called REMTM. In Yang and Li (2005) and the Chapter 4, an augmented Gibbs sampler was developed to estimate parameters of the D-K selection models. These Gibbs samplers along with the Data Augmentation of Schafer (1997; see Chapter 3) can be all used to make imputations for intermittent missing values or dropouts. For example, the hybrid Gibbs sampler in Chapter 6 can be used directly to make model-based imputations. Running the Gibbs sampler, we obtain \((\psi^{(0)}, Y_{mis}^{(0)}, ..., \psi^{(T)}, Y_{mis}^{(T)})\), from which a subset of \((Y_{mis}^{(T_0)}, ..., Y_{mis}^{(T)})\) can be selected as multiple imputations \((T_0\) represents the burning period).

Again, depending on the model used for analysis, various imputation schemes can be realized by the specific version of the Gibbs samplers. For imputation based on the pattern-mixture model with restriction identification, the procedure of Thijs et al (2002) can be used to draw imputations. For technical details, refer to the data augmentation of Schafer (1997).

If there is empirical evidence suggesting an ignorable mechanism for intermittent missingness, the data augmentation can be applied directly. More specifically, the PROC MI of SAS for close-to-monotone missingness patterns can be applied. Otherwise, the model-based hybrid Gibbs sampler can be used to make partial imputations in MPI. A simple way is to keep only the imputed \(Y_{DM}^{IM}\) after obtaining \(m\) imputations for \(Y_{mis}^{mix} = (Y_{DM}^{IM}, Y_{DM}^{DM})\).

The 2-stage MPI is mainly used for conducting sensitivity analysis. By sequentially applying a Gibbs sampler for imputing \(Y_{DM}^{IM}\) and a potentially different Gibbs sampler for imputing \(Y_{DM}^{DM}\), we can obtain multiple complete data sets that incorporate various assumptions on the mechanism of missingness and dropout.

5.5 Application of P-M to the Smoking Cessation Study

For the carbon monoxide data, we applied the P-M modeling approach to further investigate on the treatment efficacy of the two behavioral therapies and their interaction with dropout patterns. The two strategies (identifying models and
identifying restrictions) were applied within the framework of MPI and the 2-stage MPI.

5.5.1. Adding Dropout Indicator as Covariate

The CO-level data has 36 repeated measures. Even after applying partial imputation, the maximum number of dropout patterns is still too large (see Table 1.1) given the sample size of 174. A simple but meaningful way is to divide subjects into two strata depending on whether they dropped out or not. Therefore, we fitted a random effect P-M model similar to the one in (5.4), where indicator variable DROP has the same definition (DROP=1 if dropout, =0 otherwise).

In applying P-M models, graphical plots such as the ones in Figures 5.1 and 5.2 are very helpful. In Figure 5.1, the mean CO levels are drawn during the 12-week study for 62 participants who withdraw immaturely (i.e. dropouts) and 112 participants who have completed the study (i.e., completers). From the plot in Figure 5.1, it is clearly seen that the dropouts have higher average CO levels than completers especially before the 8th week. After the 8th week, the two mean curves approximately overlap with each other, which may lead to the misinterpretation that treatment effects become equal for the completers and the dropouts. An explanation is that many dropouts have already withdrawn from the study during the first 8 weeks and those who withdrew later would show similar performance with the completers. This is a natural phenomenon observed in many addiction clinical trials: the better treatment effect the participants experience, the longer they will remain in the study.

![Figure 5.1 Repeated mean CO levels across for dropouts (i.e., dropout=1) and completers (i.e. dropout=0)](image-url)
Let us first test whether there is big difference between dropouts and completers. The following random effects model was used for this purpose:

\[
y_{ij} = \beta_0 + \beta_1 T_{i1} + \beta_2 T_{i2} + \beta_3 (T_{i1} \times DROPT) + \beta_4 BaseCO + \\
+ \gamma_{0i} + \gamma_{1i} t_{ij} + \varepsilon_{ij}
\]  

(5.13)

where \( T_{i1} \) and \( T_{i2} \) have same definition as \( \text{Time1} \) and \( \text{Time2} \) in model 4.18, and \( BaseCO \) represents the baseline CO levels. We used MIXED of SAS (or the mixed-model function of MPI 2.0) to fit the mixed model and the estimates for fixed effects are shown below, from which we see significant difference between dropouts and completers measured by \( \hat{\beta}_3 \) (p=0.0005).

| Effect      | Estimate | Error | DF  | t Value | Pr > |t| |
|-------------|----------|-------|-----|---------|-------|
| Intercept   | 12.7963  | 1.0284| 172 | 12.44   | <.0001|
| T1          | -2.3667  | 0.2363| 4818| -10.01  | <.0001|
| T2          | 0.02736  | 0.00632| 4818| 4.12    | <.0001|
| T1*DROP     | -0.8444  | 0.2409| 4818| -3.50   | 0.0005|
| BaseCO      | 0.2687   | 0.0451| 4818| 5.96    | <.0001|

Our main interest for this analysis is in estimating the treatment effect of contingency management (i.e., CMs vs No-CMs). Figure 5.2 shows the mean curves for sub-groups stratified by treatment conditions (CMs vs. No-CMs) and dropout patterns (dropout vs. complete). From this plot, we can observe the following trends: (1) mean CO levels are generally higher for the dropouts than for completers; (2) within each dropout pattern (dropouts or completers) the mean CO levels are higher for No-CMs group than CMs groups. Compared with the numbers of completers within CMs group and within No-CMs group, the number of dropouts within CMs group and N-CMs group are relatively small, thus corresponding to higher variances of the CO levels. After a stepwise model selection procedure, we fitted the following model as the best random effects P-M model for this data set,

\[
y_{ij} = \beta_0 + \beta_1 T_{i1} + \beta_2 T_{i2} + \beta_3 (T_{i1} \times CM_i) + \beta_4 BaseCO + \\
+ \beta_5 (T_{i1} \times DROPT) + \beta_6 (T_{i1} \times CM_i \times DROPT) + \\
+ \gamma_{0i} + \gamma_{1i} t_{ij} + \varepsilon_{ij}
\]  

(5.14)

where predictors have similar definition to the model in (5.13) except for the treatment condition indicator \( CM_i \) (\( CM_i =1 \) if subject \( i \) received contingency management; \( CM_i =0 \) otherwise). Using Proc MIXED of SAS or MPI 2.0, the following estimates for fixed effects are obtained.

| Effect   | Estimate | Error   | DF  | t Value | Pr > |t| |
|----------|----------|---------|-----|---------|-------|
| Intercept| 12.4740  | 1.0073  | 172 | 12.38   | <.0001|
| T1       | -2.0345  | 0.3086  | 4816| -6.59   | <.0001|
| T2       | 0.02738  | 0.00632| 4816| 4.31    | <.0001|
| T1*CM    | -0.6233  | 0.3964  | 4816| -1.57   | 0.1160|
| BaseCO   | 0.2845   | 0.04395| 4816| 6.47    | <.0001|
| T1*DROP  | 0.7523   | 0.3388  | 4816| 2.22    | 0.0264|
This model partially confirms the results given by the model in (5.13), i.e. DROP is a significant predictor of CO levels. The estimate of $\beta_3$ (i.e., the effect of T1*CM) is $-0.6233$, which supports that CMs (interventions involving contingency management) outperform No-CMs (interventions without contingency management). Although the corresponding t-statistics $T_{4816}=-1.57$ comes with a p-value of 0.116, a type-3 F-test still support that T1*CM is a significant factor in model selection ($F_{1,4816}=8.92$ with p-value=$0.0028$). Note that this test on $\hat{\beta}_3 = 0$ is made after adjusting for the effects of DROP and its interaction with other covariates (i.e., T1 and BaseCO).

![Figure 5.2](image)

**Figure 5.2** Mean CO levels for completers (solid lines) and dropouts (dashed lines) in the CM (red lines) and No-CM groups (blue lines)

Additionally, we tested the treatment effects of contingency management for only completers or dropouts. We found very consistent results in the same direction favoring the contingency management. The significance level measure by P-values varies between 0.0003 (using completers only) and 0.20 (using dropouts only). Considering that the number of dropouts in this study is much smaller than the number of completers, we probably have a concrete ground to trust our analysis results, which are shown by model 5.14.

From the above analyses, we have seen that participants who dropped out of study very late show similar urinal results. This suggests another way of stratification on dropout patterns. By redefining DROP as: “DROP equates to 1 if participants withdrew before the end of 7th week, or 0 if otherwise,” we might have more meaningful interpretation. Using this version of DROP, we repeated the above
analyses, but found consistent results except that the T1*CM effect becomes less significant (T_{2977} = -1.31 with p-value=0.1917).

5.5.2 P-M Models via Averaging over Patterns

Instead of treating dropout pattern as a covariate, we fitted linear mixed models separately within each of the two groups (completers vs. dropouts). Within each group, the efficacy of CM was investigated and then averaged to represent the overall treatment efficacy. As seen from Figure 5.2, contingency management seems to be less effective for the withdrawers. Similar to the selection model application (see Chapter 4), the mixed model with AR(1) covariance structure and predictors, CM, BaseCO, and Patches, was selected for analyzing the carbon monoxide levels starting from the second week,

\[ y_{ij} = \beta_0 + \beta_1 CM_i + \beta_2 BaseCO_i + \beta_3 Patches_i. \]  

Here, we excluded the first week data from consideration to make the analysis easier.

This model was applied separately for the completers and withdrawers. Let \( \hat{\beta}_c \) and \( \hat{\beta}_w \) denote the point estimators of \( \beta_i \) respectively for the completers and withdrawers, and \( \hat{\pi}_c = 112/174 = 64\% \) the estimated probability of being completion, then the overall pointer estimator across dropout patterns is

\[ \hat{\beta}_i = \hat{\pi}_c \hat{\beta}_c + (1 - \hat{\pi}_c) \hat{\beta}_w. \]  

The corresponding variance is derived using the delta method, i.e.,

\[ \text{Var}(\hat{\beta}_i) = \frac{\partial}{\partial \beta_i} \left[ \hat{\beta}_c, \hat{\beta}_w \right] \left[ \hat{\beta}_c, \hat{\beta}_w \right]^T \hat{V}(\hat{\beta}_c) \left[ \hat{\beta}_c, \hat{\beta}_w \right] \frac{\partial}{\partial \hat{\beta}_c} \hat{\pi}_c + \hat{\beta}_w, \]  

where \( \hat{V}(\hat{\beta}_c) \) and \( \hat{V}(\hat{\beta}_w) \) represent the variances of \( \hat{\beta}_c \) and \( \hat{\beta}_w \), respectively,

\[ \hat{V}(\hat{\pi}_c) = \hat{\pi}_c (1 - \hat{\pi}_c), \quad \frac{\partial}{\partial \hat{\beta}_c} = \hat{\pi}_c, \quad \frac{\partial}{\partial \hat{\beta}_w} = \hat{\pi}_w, \]  

and \( \frac{\partial}{\partial \hat{\pi}_c} = (\hat{\beta}_c - \hat{\beta}_w). \)

Since the rates of missing information due to intermittent missing were low, only 3 instead of 4 partial imputations were created for the intermittent missing values in this analysis using pattern-mixture model. The pattern-averaged point estimators and standard errors for the treatment effect of \( \beta_i \) are listed in Table 5.1, which were further consolidated using Rubin’s rule to get the overall mean
\( \bar{\beta}_1 = -0.25 \) with standard error \( \sqrt{\text{Var}(\bar{\beta})} = 0.13 \). The test based on the t-test suggests a p-value of 0.06.

Instead of a mixed model, the D-K selection model was also used to analyze the carbon monoxide levels of the withdrawers. After averaging across patterns and imputations, we obtained consistent result with respective to the treatment effect, but there was no enough evidence in supporting a non-ignorable outcome-dependent dropout mechanism for the withdrawers (see details in Yang et al., 2005).

<table>
<thead>
<tr>
<th>Imputations</th>
<th>Completers</th>
<th>Withdrawers</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.35 (0.06)</td>
<td>-0.11 (0.09)</td>
<td>0.26 (0.13)</td>
</tr>
<tr>
<td>2</td>
<td>-0.34 (0.05)</td>
<td>-0.07 (0.10)</td>
<td>0.24 (0.13)</td>
</tr>
<tr>
<td>3</td>
<td>-0.34 (0.06)</td>
<td>-0.10 (0.09)</td>
<td>0.25 (0.13)</td>
</tr>
</tbody>
</table>

**Table 5.1** Estimated Treatment Effect of Contingency Management (\( \bar{\beta}_1 \) (S.D.)) using the Pattern-Mixture Model with two Patterns.

### 5.5.3 P-M Models with Identifying Restrictions within the 2-Satge MPI

When the number of target dropout patterns becomes large, the application of pattern-mixture models without imputation would be less affordable. For example, the mean profiles of carbon monoxide levels across four picked dropout patterns are plotted in Figure 5.3, from which we observe notable variances across and within the four patterns. As the pattern number goes up, the numbers of pattern-specific subjects become smaller, and it would be tedious or infeasible to conduct pattern-specific analysis and then combine the results across patterns. Hence, the approach of 2-stage MPI with restriction identification provides a natural alternative solution.

Adopting the procedure described in Thijs (2002), three restriction schemes (CCMV, NCMV, and ACMV) were used to impute the dropouts. Within this 2-stage MPI, the numbers of partial imputations were set as \( m = 2 \) for the first stage and \( n = 3 \) for the second stage. So together, six complete data sets were generated.
For each of them, the AR(1) mixed model with predictors \(CM\), \(BaseCO\), and \(Patches\) was applied. Using the consolidation procedure as described in Sections 5.4, the overall point estimates and fractions of missing information for the treatment effect of CM are shown in Table 5.2 along with the p-values of a one-sided hypothesis test using the t-statistics. It is seen that the overall fraction of missing information due to intermittent and dropout is much higher than that due to intermittent missingness alone. Two out of three identification strategies strongly support the favorable treatment efficacy of CM.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Overall Estimate (S.D)</th>
<th>FMI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCMV</td>
<td>-0.46 (0.22)</td>
<td>11%</td>
<td>0.02</td>
</tr>
<tr>
<td>ACMV</td>
<td>-0.42 (0.19)</td>
<td>9%</td>
<td>0.01</td>
</tr>
<tr>
<td>NCMV</td>
<td>-0.43 (0.28)</td>
<td>16%</td>
<td>0.06</td>
</tr>
</tbody>
</table>

**Figure 5.3** Pattern-Dependent Mean Carbon Monoxide Levels

**Table 5.2** Estimated Treatment Effect of Contingency Management using the Pattern-Mixture Models within the Framework of 2-stage MPI
The idea of three restriction schemes is depicted in Figure 5.4, which shows the dropout patterns for a data set with 6 repeated measures and 5 dropout patterns. To draw imputations, the missing values in the white blocks are imputed in the order as numbered. That is, we first impute missing values on $Y_2$ in Pattern 5, followed by imputing missing values on $Y_3$ in Pattern 4, and so on. Using the missing values on $Y_4$ in Pattern 3 as an example, CCMV corresponding to “$W_{13} = 1$ and $W_{23} = 0$.” This means that the conditional distribution $f(Y_4 | Y_1, Y_2, Y_3)$ in Pattern 1 (i.e., completers) together with the observed values on $(Y_1, Y_2, Y_3)$ in Pattern 3 is used to predict the value of $Y_4$ in Pattern 3. The restriction scheme NCMV corresponds to “$W_{13} = 0$ and $W_{23} = 1$,” where information from the neighbors (in Pattern 2) are borrowed. Similarly, we can specify “$W_{13} = W_{23} = 0.5$” to represent the restriction scheme ACMV. It is also flexible to set other values for the weights to represent different ways of restriction, e.g., “$W_{13} = 0.6$ and $W_{23} = 0.4$,” which can be viewed as a mixture of CCMV and NCMV.

**Figure 5.4** Identification of Restrictions in the 2-Stage MPI

<table>
<thead>
<tr>
<th></th>
<th>Y1</th>
<th>Y2</th>
<th>Y3</th>
<th>Y4</th>
<th>Y5</th>
<th>Y6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pattern 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>W11</td>
<td>W12</td>
<td>W13</td>
<td>W14</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Pattern 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>W21</td>
<td>W22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Pattern 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>W31</td>
<td>W32</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Pattern 4</strong></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>W41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Pattern 5</strong></td>
<td></td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

**SUMMARY FOR CHAPTER 5**

In this chapter, another modeling approach called P-M models was introduced. When implementing this group of models, two strategies were summarized. We strongly advocated the application of the method with identification of restrictions. The MPI and 2-stage MPI were proposed here, which jointly provide
the frameworks for sensitivity analysis. As seen in the following previous and the
following chapters, various methods exist in conducting sensitivity analysis.

MPI 2.0 has implemented the 2-stage MPI using CCMV, ACMV, and NCMV by
specifying the weight matrix as seen in Figure 5.4. The MCMC algorithms for D-
K (Yang and Li, 2005) and REMTM (see Chapter 6) can be chosen for making
partial imputations for intermittent missing values and dropouts. The smoking
cessation CO-level data was reanalyzed using P-M models with different
strategies and schemes of restrictions. All the analyses jointly support the
favorable treatment effect of contingency management.

For more details on theoretical discussions, please refer to Yang et al. (2005) and
references thereafter.
CHAPTER 6.

RANDOM-EFFECTS MARKOV TRANSITION MODELS FOR NONIGNORABLE INTERMITTENT MISSINGNESS & DROPOUT

In the earlier phase of our research on incomplete longitudinal modeling, the D-K selection and P-M models were emphasized. Recently, our group realized the promising potential of a new group of modeling approach called Shared-Parameter (S-P) models. For nonignorable dropouts, D-K and P-M models are convenient choices. But for nonignorable intermittent missing values, the S-P models can be more easily applied. Here in this chapter, a specific technique called Random-Effects Markov Transition Model (REMTM) is introduced. For the illustration purposes, REMTMs for binary repeated measures are described, although similar models for count and continuous repeated measures are also available in MPI 2.0. As a recap in summarizing the available choices for nonignorable missing values, the concept of ignorability within three settings of modeling is re-conceptualized first.

6.1 Incomplete Longitudinal Data Analysis

6.1.1 Selection, Pattern-Mixture, and Shared-Parameter Models

For a longitudinal data set with balanced design, \( J \) repeated measures are potentially observed on each of the \( N \) subjects at times \( t_{i1}, ..., t_{iJ} \) \((i = 1, ..., N; j = 1, ..., J)\). For the following discussion, we use capital symbols to represent variables, e.g., \( Y_1, ..., Y_J \) indicating response variables, and \( X_1, ..., X_K \) indicating covariates or explanatory variables. Symbols in lower case represent observed or missing values: \( y_{ij} \) denoting the value of \( Y_j \) and \( x_{ijk} \) denotes the value of \( X_k \) recorded at time \( t_{ij} \) \((i = 1, ..., N; j = 1, ..., J; k = 1, ..., K)\). Bold symbols represent vectors or matrices, e.g., the vector \( y_i = (y_{i1}, ..., y_{ij})^T \) indicating values of repeated measures and the matrix \( X_i = [x_{ijk}]_{J \times K} \) consisting of values of time-varying or -independent covariates for the \( i^{th} \) subject. Assuming that repeated measures are distributed as multivariate normal, a repeated-measures model with structured covariance matrix can be written as \( y_i = X_i \beta + \epsilon_i \) where \( \epsilon_i \sim N(0, \Sigma_i) \) and \( \beta \) is a vector of fixed-effects parameters. Determined by the way of parameterization of the covariance matrix, various forms of mixed models can be derived (Jennrich and Schluchter, 1986).
When some values of repeated measures are missing, we partition $y_i$ into two parts, $y_i = (y_{i}^{obs}, y_{i}^{mis})$, with $y_{i}^{obs}$ indicating the observed values, and $y_{i}^{mis}$ indicating values that would be observed if they were not missing. A vector of missingness indicators is defined as $r_i = (r_{i1}, \ldots, r_{iT})^T$ with elements $r_{ij} = 0$ (or 1) indicating whether $y_{ij}$ is observed (or missing). Theoretically, the joint distribution of the observed data (i.e., $y_{i}^{obs}$) and missingness patterns (i.e., $r_i$) should be modeled in statistical analysis based on the full likelihood function,

$$L(\theta, \phi \mid y_{i}^{obs}, r_i) \propto \prod_{i=1}^{N} \int f(y_i, r_i \mid \theta, \phi) dy_{i}^{mis} \quad (6.1)$$

where $\theta$ represents parameters of the model for repeated measures, and $\phi$ represents parameters of the missingness mechanism. According to the possible causal path, there exist three ways in factoring the joint distribution of the complete data and missingness indicators: outcome-dependent factorization, pattern-dependent factorization, and parameter-dependent factorization (see Figure 6.1). Accordingly, there are three models available for incomplete longitudinal data analysis.

(A) Selection Model, which factors the joint distribution $f(y_i, r_i \mid X_i, \theta, \phi)$ into a marginal distribution for $y_i$ and a conditional distribution of $r_i$ given $y_i$, i.e.,

$$f(y_i, r_i \mid X_i, \theta, \phi) = f(y_i \mid X_i, \theta) f(r_i \mid y_i, X_i, \phi) \quad (6.2)$$

where $f(r_i \mid y_i, X_i, \phi)$ can be interpreted as “self-selection of the $i^{th}$ subject into a specific missingness group.” This outcome-dependent missingness model was first developed by Diggle and Kenward (1994), where a logistic regression model is used to model the dropout probability in studying protein content of cow milk.

(B) Pattern-Mixture Model, which is a pattern-dependent model assuming that distribution of repeated measures varies with the missingness patterns and the joint distribution is factored as

$$f(y_i, r_i \mid X_i, \theta, \phi) = f(y_i \mid r_i, X_i, \theta) f(r_i \mid X_i, \phi). \quad (6.3)$$

Assuming that there are $P$ patterns of missingness in a data set, the marginal model of $y_i$ is a mixture model,

$$f(y_i) = \sum_{p=1}^{P} f(y_i \mid r_i = p, X_i, \theta^{(p)}) \pi_p \quad (6.4)$$
where $\theta^{(p)}$ represents the parameters of $f(y_i)$ in the $p^{th}$ pattern, $\pi_p = \Pr(r_i = p \mid X_i, \phi)$, and $r_i$ is replaced by a scalar $r_i$ to numerate the $P$ patterns ($p = 1, \ldots, P$). It should be emphasized here that $\theta^{(1)}, \ldots, $ and $\theta^{(P)}$ may be different both in dimensionality and in value.

(C) **Shared-Parameter Model**, which assumes that $y_i$ and $r_i$ are conditional independent of each other, given a group of parameters $\xi_i$, i.e.,

$$f(y_i, r_i \mid X_i, \theta, \phi) = \int f(y_i \mid \xi_i, X_i, \theta) f(r_i \mid \xi_i, X_i, \phi) f(\xi_i) \, d\xi_i$$  \hspace{1cm} (6.5)

From the point view of causation, shared “parameters” $\xi_i$ play the role of a confounder for the relationship between $y_i$ and $r_i$, thus can be either observable variables (e.g., gender) or latent variables (e.g., random-effects). For the case of observed confounders, model (6.5) is in fact a mixture model and the analysis can be conducted by a stratification analysis, e.g., divide subjects into two strata according to their gender and within each stratum $y_i$ and $r_i$ are modeled independently.

**Figure 6.1** Three Ways of Defining Missingness Mechanism

### 6.1.2 Ignorable versus Nonignorable Missingness

In certain biomedical studies, both missingness patterns and values of repeated measures are of interest. For example, in a heart-disease study, the repeatedly measured blood pressures and the survival lengths of the patients should be modeled jointly. In these scenarios, the above selection, pattern-mixture, and shared-parameter models can be applied directly or after some modification (Hogan and Laird, 1997). In a majority biomedical studies, however, only the
parameters of repeated measures themselves are of interest, while parameters related to missing values are usually viewed as nuisance. In this latter case, it is desirable that missing data be ignored.

Within the setting of outcome-dependent missingness, the concept of “ignorability” was defined and extensively addressed. According to Rubin (1976), missing values can be ignored only when two conditions are satisfied: (i) $r_i$ is independent of $y_{i}^{mis}$, given $y_{i}^{obs}$ and $X_i$; (ii) $\theta$ and $\phi$ are distinct. Under this ignorability, the likelihood function for $\theta$ can be separated from the likelihood function for $\phi$, i.e.,

$$l(\theta,\phi \mid y_{i}^{obs}, r_i) = l(\theta \mid y_{i}^{obs}) + l(\phi \mid y_{i}^{obs}, r_i)$$ (6.6)

where $l(\cdot) = \log(L(\cdot))$ indicates log-likelihood function. In Little and Rubin (2002), outcome-dependent missingness was further divided into sub-categories: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). It is clear now that most longitudinal models based on observed-data likelihood (i.e., $l(\theta \mid y_{i}^{obs})$) require that missing data are ignorable, e.g., marginal models with generalized estimating equations requiring the assumption of MCAR (Robins et al., 1995) while generalized linear mixed models and Markov transition models assume MAR (Hall et al., 2001). In practice, without computational tools specially designed for nonignorable missing values, the assumptions of ignorability is the default assumption in many software packages.

**Figure 6.2** Intermittent Missing Values and Dropouts in an Imaginary Clinical Trial Data with Two Subjects
For intermittent missing values, ignorability can be interpreted as whether they could be interpolated from neighborhood observed values. For dropouts, the assumption corresponds to whether missing values after dropout can be extrapolated from the previous observed values. In certain applications, occasional omission or nonresponses are usually due to reasons that are purely of randomness in nature, e.g., schedule conflicts or bad weather. Thus, intermittent missing values can be assumed to be ignorable. Nonetheless, subjects withdraw from a study usually because of study-related reasons, e.g., being unsatisfactory with the intervention or notorious side effects of a medical therapy, hence dropouts are usually nonignorable (Murray and Findlay, 1988; Diggle et al., 2004; and Verbeke and Molenberghs, 2000). Figure 6.2 depicts an imaginary clinical trial data with two subjects: one receiving medicine, one receiving placebo. A common strategy in the pharmaceutical data analysis is “last observation carried forward.” From this plot, it is clearly seen that this method would lead to biased results in this example data, which represents a typical scenario of practical settings.

The definition of ignorability should be extended to meet the needs of pattern-mixture and shared-parameter models. Adopting an informal way, this article redefines “ignorability” as a condition under which observed data can be used to estimate \( \theta \) without bias. For selection and pattern-mixture models, so long as \( r_i \) and \( y_i^{mis} \) are independent of each other, given \( y_i^{obs} \) and \( X_i \), missing data can be ignored. For shared-parameter models, ignorability corresponds only to the case where \( \xi_i \) are observable confounders, which are usually viewed as a subset of \( X_i \). Unless \( r_i \) and \( y_i \) share no random-effects, a shared-parameter model would generally associate with a nonignorability assumption.

### 6.2 REMTMs for Nonignorable Intermittent Missingness and Dropout

When the dynamic features of the transition pattern in longitudinal data are of interest, an appropriate longitudinal approach is a Markov transition model. For binary repeated measures with nonigorable missing values, Albert and Follmann (2003) developed a Markov transition model with random-effects that were shared by the sub-model on measurement and the sub-model on missingness indicators.

In the REMTM for incomplete binary repeated measures, the sub-model for measurement process assumes a first-order Markov chain for each series of binary measures. The transitional probabilities \( P_{kl} = f(y_{ij} = l | y_{i,j-1} = k) \) \((k = 0 \text{ or } 1; l = 0 \text{ or } 1)\) can be modeled by a logistic regression with random intercepts,
\[ \text{logit}(P_{01}) = \text{logit}(y_{ij} = 1 \mid y_{i,j-1} = 0, x_{ij}, \xi_i) = x_{ij}'B_{01} + \xi_i, \]
\[ \text{logit}(P_{10}) = \text{logit}(y_{ij} = 0 \mid y_{i,j-1} = 1, x_{ij}, \xi_i) = x_{ij}'B_{10} + \nu_{ij}, \]

(6.7)

where \( \xi_i \sim N(0, \sigma_{\xi}^2) \) denotes the random intercept, and \( \nu \) is the heterogeneity parameter indicating the correlation between \( P_{10} \) and \( P_{01} \).

The distribution of missingness indicators \( r_i = (r_{i1}, \ldots, r_{iL})' \) can be modeled by another Markov transition model. Redefining

\[
r_{ij} = \begin{cases} 
0 & \text{if } y_{ij} \text{ is observed} \\
1 & \text{if } y_{ij} \text{ is missing intermittently} \\
2 & \text{if } y_{ij} \text{ is missing due to dropout}
\end{cases}
\]

(6.8)

a first-order Markov process associates with \( 3 \times 3 \) transition probabilities (i.e., \( P_{kl} = \Pr(r_{ij} = l \mid r_{i,j-1} = k) \): \( k = 0, 1, 2; \ l = 0, 1, 2 \)). Determined by certain restrictions, the following transition probabilities would be always equal to zero: \( P_{12} = P_{20} = P_{21} = 0 \). For other combinations of \( r_{i,j-1} \) and \( r_{ij} \), the transition probabilities are calculated in the following way. First, if the previous count measure is observed (i.e., \( r_{i,j-1} = 0 \)), then the current one could be observed, intermittently missing, or dropout, and the 3-category multinomial-logit model (Agresti, 2002) can be used to calculate the transition probabilities,

\[
P(r_{ij} = k \mid \xi_i, x_{ij}, r_{i,j-1} = 0) = \begin{cases} 
1 & \text{if } k = 0, \\
\frac{1 + \sum_{l=1}^{2} \exp(x_{ij}'\eta_l + \xi_i'\gamma_l)}{1 + \sum_{l=1}^{2} \exp(x_{ij}'\eta_l + \xi_i'\gamma_l)} & \text{if } k = 1 \text{ or } 2.
\end{cases}
\]

(6.9)

Second, if the previous measure is intermittently missing, then the current one may only be observed or intermittently missing again. Correspondingly, a logistic regression model is used to calculate \( P_{10} \) and \( P_{11} \), i.e.,

\[
P(r_{ij} = k \mid \xi_i, x_{ij}, r_{i,j-1} = 1) = \begin{cases} 
1 & \text{if } k = 0, \\
\frac{1 + \exp(x_{ij}'\eta_l + \xi_i'\gamma_l)}{1 + \exp(x_{ij}'\eta_l + \xi_i'\gamma_l)} & \text{if } k = 1.
\end{cases}
\]

(6.10)

Third, for the absorbing state, we would always have \( P(r_{ij} = 2 \mid \xi_i, x_{ij}, r_{i,j-1} = 2) = 1 \). Denoting \( T_i \) as the time for the last observed measurement for subject \( i \), special considerations should also be given to \( y_{iT_i} \), for
which we always have \( P(r_{i,t} = 0 | r_{i,t-1} = 1) = 1 \). In the above logit and logistic regression models, regression coefficients \( \eta_1 \) and \( \eta_2 \), respectively, indicate whether intermittent missingness and dropout depend on covariates, while \( \gamma_1 \) and \( \gamma_2 \) respectively indicated whether the two types of missing values are nonignorable.

By combining the above sub-models for measurement and missingness, the likelihood function for parameters \( \theta = (\beta_{01}, \beta_{10}, \nu, \sigma^2_\xi) \) and \( \phi = (\eta_1, \eta_2, \gamma_1, \gamma_2)^T \) is expressed as

\[
L(\theta, \phi) \propto \left( \prod_{i=1}^N \left( \prod_{j=1}^{T_i} p(y_{ij} | x_{ij}, y_{i,j-1}, \xi_i, \theta) \right) \right) \left( \prod_{j=1}^{T_i} p(r_{ij} | x_{ij}, r_{i,j-1}, \xi_i, \phi) \right) p(\xi_i) d\xi_i . \quad (6.11)
\]

It should be remarked here that the above REMTM can be easily extended to deal with other types of repeated measures. Li et al. (2005) applied the REMTMs to Poisson-distributed repeated measures with nonignorable missing values. The random-intercept in (6.7)-(10) can be replaced with other types of random effects such as random slopes and random cohort effects. The REMTM is only one specific example of shared-parameter models, other longitudinal models such as marginal model or random-effects models can be also used to implement shared-parameters modeling. Actually, shared-parameter model was first developed by Wu and Caroll (1988) where certain parameters are shared by the measurement model and a censoring process. Other examples of shared-parameter models are seen in Little (1995), Wu and Bailey (1998), Wu and Follmann (1999), Albert (2000), Albert and Follmann (2002), Follman and Wu (1995), Pulkstenis, et al. (1998), and Ten Have et al. (1998).

### 6.3 MCMC Algorithms for Model Fitting and Imputation

#### 6.3.1 A Hybrid Gibbs Sampler for Fitting REMTM

For the REMTM, we denote \( \psi = (\theta, \phi, \xi_i, \sigma^2_\xi) \) and \( y_i = (y^{obs}_i, y^{mix}_i) \) without differentiating intermittent missing values from dropouts. By setting “\( t = 0 \)” and initializing the parameters and missing values with \( \psi = \psi^{(0)} \) and \( y^{mix}_i = y^{mix(0)}_i \), we repeat the following Gibbs steps.
1. **Imputation Step**: Draw imputations for missing values. For \( i = 1, \ldots, n \) and \( j = 1, \ldots, J \), if \( y_{ij} \) is missing, an imputation would be drawn using
\[
y_{ij}^{\text{mis}(t+1)} \sim f(y_{ij} | y_{i,j-1}^{(t)}, y_{i,j+1}^{(t)}, \psi^{(t)})
\]
where \( y_{i,j-1} \) and \( y_{i,j+1} \) are observed or imputed values at the previous iteration.

2. **Estimation Step**: draw parameters and random effects in the following order,
\[
\begin{align*}
\theta^{(t+1)} & \propto f(\theta) \prod_{i=1}^{n} f(\theta | y_i^{\text{obs}}, y_i^{\text{mis}(t+1)}, \varphi^{(t)}, \xi_i^{(t)}, \sigma^2_{\xi}^{(t)}) \\
\varphi^{(t+1)} & \propto f(\varphi) \prod_{i=1}^{n} f(\varphi | y_i^{\text{obs}}, y_i^{\text{mis}(t+1)}, \theta^{(t+1)}, \xi_i^{(t)}, \sigma^2_{\xi}^{(t)}) \\
\xi_i^{(t+1)} & \propto f(\xi_i) \prod_{i=1}^{n} f(\xi_i | y_i^{\text{obs}}, y_i^{\text{mis}(t+1)}, \theta^{(t+1)}, \varphi^{(t+1)}, \sigma^2_{\xi}^{(t)}) \\
\sigma^2_{\xi}^{(t+1)} & \propto f(\sigma^2_{\xi}) \prod_{i=1}^{n} f(\sigma^2_{\xi} | y_i^{\text{obs}}, y_i^{\text{mis}(t+1)}, \theta^{(t+1)}, \varphi^{(t+1)}, \xi_i^{(t+1)})
\end{align*}
\]
where \( f(\theta), f(\varphi), f(\xi_i), \) and \( f(\sigma^2_{\xi}) \) are prior distributions of the parameters, and \( y_i^{\text{mis}(t+1)} \) represents the current imputed values.

3. Set \( t = t + 1 \) and go to step 1.

Running this Gibbs sampler with large enough iterations, the procedure would converge under regularity conditions and we obtain a series \( \psi^{(0)}, y_i^{\text{mis}(0)}, \ldots, \psi^{(T)}, y_i^{\text{mis}(T)} \). By discarding the first \( T_0 \) burning samples (e.g., \( T_0 = 10\% \times T \)), \( (\psi^{(T_0)}, \ldots, \psi^{(T)}) \) can be used to estimate the posterior distribution of \( \psi \) and further inferences can be made accordingly. This algorithm provides a structure for fitting any forms of REMTM depending on the types of repeated measure. For Poisson-distributed count data, see implementations in Li et al., (2005).

Depending on the option chosen for modeling measurements and missingness mechanisms, different version of Gibbs samplers are conceivable. For example, the selection model for continuous repeated measures with nonignorable dropouts can be implemented with \( \psi = (\alpha, \beta, \varphi) \), a Gaussian distribution (i.e., \( f(y_i^{\text{mis}} | y_i^{\text{obs}}, \psi) \)) for drawing missing values, and logistic regression modeling the dropouts (Yang and Li, 2005).

### 6.3.2 Sampling Conditional Distributions and Prior Specification
The above algorithm is called hybrid Gibbs sampler because various sampling schemes can be embedded to simulate parameters from the fully conditional densities. If a conditional distribution has a known form, the corresponding parameter vector is sampled directly. For example, missing values of continuous repeated measures in selection model can be sampled from a Gaussian distribution. If the conditional distribution has a log-concave form after proper transformation (e.g., the density for regression coefficients and residual variance), the efficient scheme called adaptive rejection sampling (Gilks and Wild, 1992) can be applied. Otherwise, the less efficient but more robust methods can be applied, such as Metropolis-Hasting or the giddy Gibbs sampler (Ritter and Tanner, 1992).

When there is no historical data or auxiliary information at hand, a convenient choice is the flat or non-informative prior. According to our experience, flat uniform distributions for covariance parameters or diffused normal distributions for regression coefficients usually work well. In certain situation, conjugate priors should be adopted if possible, e.g., a Normal-Wishart distribution for the joint distribution of mean and covariance matrix of a multivariate normal repeated measures.

For details on Gibbs sampler implementation, please see Yang and Li (2005). MPI 2.0 has implemented several MCMC algorithms for REMTM for binary, count, and continuous repeated measures.

6.4 Application

6.4.1 REMTMs for Dichotomized Carbon Monoxide Data

For the same group of carbon monoxide levels, we reanalyzed them using the REMTM after a dichotomization to indicate use or abstinence of cigarette. This dichotomized version of carbon monoxide data was analyzed by Yang et al. (2005) using a REMTM with maximum likelihood estimation. Here it was reanalyzed using the hybrid Gibbs sampler for REMTM with predictors: $CM_i$, $RP_i$, and $RP^*CM_i$ as defined in Section 6.3.1. Table 6.1 depicts the estimated posterior means, standard deviations, and 95% credible intervals (C.I.) for all the parameters of interest.

The estimated parameters for $\sigma^2$, $\gamma_1$, and $\gamma_2$ jointly suggest that both intermittent missingness and dropout are parameter-dependent nonignorable. The
introduced random intercept effects (i.e., $\xi$) capture the heterogeneity on missingness across the subjects. Among all the estimated parameters of $\eta_1$, only the one corresponding to CM is significantly different from zero (i.e., $\xi^3 = -1.19$), indicating that smokers receiving CM had smaller chance of missing the clinic visits occasionally.

The negative value of estimated $\nu$ suggests that individuals who had a large transition probability $P_{01}$ had lower transition probability $P_{10}$ (see Figure 6.3). In other words, individuals had an affinity for staying at either the “use” or “abstinence” state. Of most our interest, the fitted REMTM confirmed a strongly favorable treatment efficacy of CM by increasing the instilling probability $P_{10}$ (i.e., $\hat{\beta}^3_{10} = 2.61$ with 95% C.I.=(1.69, 3.53)) and the maintaining abstinence probability $P_{00}$ (i.e., $\hat{\beta}^3_{01} = -1.19$ with 95% C.I.=(-1.86, -0.52)).

![Figure 6.3 Transition Probabilities for the Dichotomized Carbon Monoxide Levels](image)

**Figure 6.3** Transition Probabilities for the Dichotomized Carbon Monoxide Levels
### Table 6.1 Posterior Parameter Estimation with Standard Deviation and 95% Credible Intervals Using the REMTM to the Dichotomized Carbon Monoxide Data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimates</th>
<th>Std. De.</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transition Probability ($P_{01}$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--Intercept ($\beta_{01}^1$)</td>
<td>-0.39</td>
<td>0.16</td>
<td>(-0.70, -0.07)</td>
</tr>
<tr>
<td>--RP ($\beta_{01}^2$)</td>
<td>0.14</td>
<td>0.50</td>
<td>(-0.84, 1.12)</td>
</tr>
<tr>
<td>--CM ($\beta_{01}^3$)</td>
<td>-1.19</td>
<td>0.34</td>
<td>(-1.86, -0.52)</td>
</tr>
<tr>
<td>--RP*CM ($\beta_{01}^4$)</td>
<td>0.16</td>
<td>0.61</td>
<td>(-1.04, 1.36)</td>
</tr>
<tr>
<td>Transition Probability ($P_{10}$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--Intercept ($\beta_{10}^1$)</td>
<td>-1.69</td>
<td>0.37</td>
<td>(-2.42, -0.96)</td>
</tr>
<tr>
<td>--Baclofen ($\beta_{10}^2$)</td>
<td>0.90</td>
<td>0.66</td>
<td>(0.39, 2.19)</td>
</tr>
<tr>
<td>--Baseline ($\beta_{10}^3$)</td>
<td>2.61</td>
<td>0.47</td>
<td>(1.69, 3.53)</td>
</tr>
<tr>
<td>--RP*CM ($\beta_{10}^4$)</td>
<td>-1.21</td>
<td>0.81</td>
<td>(-2.80, 0.38)</td>
</tr>
<tr>
<td>Variance of Random Intercept ($\sigma^2$)</td>
<td>3.32</td>
<td>0.64</td>
<td>(2.07, 4.57)</td>
</tr>
<tr>
<td>Heterogeneity Parameter ($\nu$)</td>
<td>-1.43</td>
<td>0.14</td>
<td>(-1.70, -1.15)</td>
</tr>
<tr>
<td>Nonignorable Missingness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--Intermittent Missingness ($\gamma_1$)</td>
<td>0.93</td>
<td>0.11</td>
<td>(0.71, 1.14)</td>
</tr>
<tr>
<td>--Dropout ($\gamma_2$)</td>
<td>0.56</td>
<td>0.14</td>
<td>(0.29, 0.83)</td>
</tr>
<tr>
<td>Covariate-Dependent Missingness for</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intermittent Missing ($\eta_1$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--Intercept ($\eta_1^1$)</td>
<td>-2.28</td>
<td>0.25</td>
<td>(-2.77, -1.79)</td>
</tr>
<tr>
<td>--RP ($\eta_1^2$)</td>
<td>-0.48</td>
<td>0.47</td>
<td>(-1.40, 0.44)</td>
</tr>
<tr>
<td>--CM ($\eta_1^3$)</td>
<td>-1.19</td>
<td>0.34</td>
<td>(-1.85, -0.52)</td>
</tr>
<tr>
<td>--RP*CM ($\eta_1^4$)</td>
<td>1.14</td>
<td>0.57</td>
<td>(0.02, 2.26)</td>
</tr>
<tr>
<td>Dropout ($\eta_2$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--Intercept ($\eta_2^1$)</td>
<td>-4.76</td>
<td>0.37</td>
<td>(-5.49, -4.03)</td>
</tr>
<tr>
<td>--RP ($\eta_2^2$)</td>
<td>0.14</td>
<td>0.52</td>
<td>(-0.88, 1.16)</td>
</tr>
<tr>
<td>--CM ($\eta_2^3$)</td>
<td>-0.21</td>
<td>0.50</td>
<td>(-1.19, 0.77)</td>
</tr>
<tr>
<td>--RP*CM ($\eta_2^4$)</td>
<td>-0.01</td>
<td>0.70</td>
<td>(-1.38, 1.36)</td>
</tr>
</tbody>
</table>
6.4.2 REMTM for the Baclofen Clinical Trial

The binary urine screening data in the Baclofen clinical trial were analyzed by Shoptaw et al. (2003) using a marginal model with generalize estimating equations and a Markov transition model. Both models jointly supported favorable treatment effects of Baclofen in reducing the probability of cocaine use or enhancing the transition probabilities instilling or maintaining abstinence of cocaine. In Figure 6.4, the marginal percentage of Benzoylecgonine(BE)-free urine samples are depicted. The stars over the x-axis indicates the time points where the p-value of the point-wise ANOVA is smaller than 0.001. Figure 6.5 shows the marginal missingness rates on each time pint across the treatment groups.

**Figure 6.4** Percentage of Benzoylecgonine(BE)-free Urine Samples in the Baclofen Trial

**Figure 6.5** Rates of Missingness for the Baclofen Study
A fundamental problem for these analyses was that missing data have been assumed of ignorable (MCAR for the marginal model and MAR for Markov transition model). Therefore, it was important to verify this assumption by fitting a REMTM.

After a model-selection process, an REMTM was fitted to the data with $x_{ij} = (1, \text{Baclofen}_i, \text{Baseline}_i)^T$, where $\text{Baclofen}_i = \text{“1” or “0”}$ respectively indicating whether the $i^{th}$ individual received baclofen or placebo, $\text{Baseline}_i$ represents the number of BE-positive urine samples during the 2-week baseline period. The parameter estimates along with 95% confidence intervals are depicted in Table 6.2 from which we may draw the following conclusions. First, compared with placebo, the beneficial efficacy of Baclofen is confirmed by seeing that it helps reducing $P_{01} (\beta^i_{01} = -1.39$ with 95% C.I.$= (-2.15, -0.63))$ and increasing $P_{10} (\beta^i_{10} = 0.97$ with 95% C.I.$= (0.06, 1.87))$. Second, estimation on $\gamma_1$ and $\gamma_2$ jointly infer that both intermittent missingness and dropout are parameter-dependent nonignorable in the sense that the missingness mechanism shares random intercept ($\xi_i$) with repeated measures. Third, there is little evidence in supporting that missingness is related to either treatment assignment or baseline severity; see estimated parameters $\eta_1$ and $\eta_2$. There are additional useful inferences provided by the model. For example, the estimate of $\nu$ indicates that $P_{01}$ and $P_{10}$ are negatively correlated, which means that participants of the study have a tendency to be either in the 0 or 1 state. Comparing the above REMTM results with those of marginal and Markov transition models, it is clearly seen that REMTM provides richer information than its peers.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimates</th>
<th>Std. Err.</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transition Probability ($P_{01}$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--Intercept ($\beta^0_{01}$)</td>
<td>0.07</td>
<td>0.42</td>
<td>(-0.89, 0.76)</td>
</tr>
<tr>
<td>--Baclofen ($\beta^1_{01}$)</td>
<td>-1.39</td>
<td>0.39</td>
<td>(-2.15, -0.63)</td>
</tr>
<tr>
<td>--Baseline ($\beta^2_{01}$)</td>
<td>0.45</td>
<td>0.12</td>
<td>(0.21, 0.69)</td>
</tr>
<tr>
<td>Transition Probability ($P_{10}$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--Intercept ($\beta^0_{10}$)</td>
<td>1.19</td>
<td>0.50</td>
<td>(0.21, 2.17)</td>
</tr>
<tr>
<td>--Baclofen ($\beta^1_{10}$)</td>
<td>0.97</td>
<td>0.46</td>
<td>(0.06, 1.87)</td>
</tr>
<tr>
<td>--Baseline ($\beta^2_{10}$)</td>
<td>-0.52</td>
<td>0.14</td>
<td>(-0.78, -0.25)</td>
</tr>
<tr>
<td>Variance of Random Intercept ($\sigma^2$)</td>
<td>2.70</td>
<td>0.40</td>
<td>(1.92, 3.48)</td>
</tr>
<tr>
<td>Heterogeneity Parameter ($\nu$)</td>
<td>-0.91</td>
<td>0.24</td>
<td>(-1.38, -0.44)</td>
</tr>
<tr>
<td>Informative Missingness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--Intermittent Missingness ($\gamma_1$)</td>
<td>0.31</td>
<td>0.06</td>
<td>(0.19, 0.42)</td>
</tr>
<tr>
<td>--Dropout ($\gamma_2$)</td>
<td>0.39</td>
<td>0.10</td>
<td>(0.18, 0.59)</td>
</tr>
</tbody>
</table>
### Table 6.2 Parameter Estimation with Standard Errors and 95% Confidence Intervals

<table>
<thead>
<tr>
<th>Covariate-Dependent Missingness for</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermittent Missing ($\eta_1$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept ($\eta_{1i}$)</td>
<td>-0.30</td>
<td>0.15</td>
</tr>
<tr>
<td>Baclofen ($\eta_{1i}$)</td>
<td>-0.03</td>
<td>0.14</td>
</tr>
<tr>
<td>Baseline ($\eta_{1i}$)</td>
<td>-0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>Dropout ($\eta_2$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept ($\eta_{2i}$)</td>
<td>-2.21</td>
<td>0.31</td>
</tr>
<tr>
<td>Baclofen ($\eta_{2i}$)</td>
<td>-0.23</td>
<td>0.29</td>
</tr>
<tr>
<td>Baseline ($\eta_{2i}$)</td>
<td>-0.09</td>
<td>0.10</td>
</tr>
</tbody>
</table>

**SUMMARY FOR CHAPTER 6**

In this chapter, the concept of ignorability was reviewed and re-interpreted within a more generalized setting involving three ways in factorizing the joint distribution of the repeated measures and missingness indicators. As a compliment to D-K selection and P-M models, S-P models were proposed and Bayesian inference based on MCMC was developed. A specific form of S-P model called REMTM was introduced and applied to two practical binary repeated measures. After a dichotomization, the CO-level data was re-analyzed using the REMTM, which supported a favorable treatment effects of contingency management. This REMTM depict the dynamic transitional features of the data and the treatment efficacy, which provide an additional way in modeling the joint distribution of multivariate repeated measures.

For illustration purposes, the REMTM for binary repeated measures was shown in this chapter, but MPI 2.0 has REMTM options to fit other type of repeated measure, too. Both simulation data and practical data (Li and Yang 2005) has proved the satisfactory performance of REMTM with MCMC fitting algorithms. Finally, we emphasize that REMTM provides a useful device in making imputations for intermittent missing values and dropouts so that the framework of MPI and 2-stage MPI can be implemented to conduct sensitivity analysis.
CHAPTER 7.

SUMMARY & DISCUSSION

This technical report mainly reviews three modeling strategies for incomplete longitudinal data using the full-likelihood functions and demonstrates their application within the MPI frameworks mainly using a carbon monoxide data set. Selection, pattern-mixture, and shared-parameter models jointly cover a wide range of longitudinal models such as marginal models, mixed-effects models. For example, the selection model degenerates to a mixed-model if the dropouts are ignorable. Because of the limit of effort, only continuous and binary repeated measures are used for illustration. In fact, all the three modeling strategies can be specified to deal with various formats of repeated measures and missingness and dropout mechanisms.

In observing that various terms are encountered in literature in describing missingness mechanism, e.g., “informative” missing or dropout (Albert and Follmann, 2003; Diggle and Kenward, 1994), “missing not at random” (Kenward and Molenberghs, 1999), and “inaccessible” missingness (Collins and Seitz, 1994), we deem it necessary to unify the terminology. When doing so, we emphasize the clear articulation between “ignorability” versus “nonignorability.” Any specific form of nonignorability should be described with appropriate adjectives. For example, MNAR would better be rephrased as “outcome-pendent nonignorable.” Although in this report, various terms in describing mechanism of missingness and dropout are still employed.

The most notable limitation with practical data analysis with missing values is that the true model and mechanism for measurements and missingness are usually unverifiable. Thus, in many settings, selection, pattern-mixture, or shared-parameter models should be viewed as models with rich assumptions. Our guideline is to always investigate the sensitivity of the inferences on fixed parameters to varying assumptions. Inspired by the idea of pattern-mixture with restriction identification, the MPI-based frameworks were proposed. In the statistical literature, model-based sensitivity analyses are sometimes seen; e.g., selection model with local influence (Verbeke and Molenburghs, 2000) and pattern-mixture mode with varying restrictions (Thijis, 2002). However, one must be careful that by stretching a specific model further to mine the data deeper would require richer and more rigorous theoretical basis that might not be supported by the practical data. This is especially true in phase I, II, and III clinical trails, where the sample sizes are usually not large enough to afford such kind of over-fitting. Adopting a more border sense, this technical report recommends the conduction of sensitivity analysis across the models within the MPI frameworks.
For the same set of data from the smoking trial, we applied various models to analyze the treatment efficacy of two behavioral therapies: contingency management and relapse prevention. Selection models and pattern-mixture models were jointly applied to the original continuous carbon monoxide levels. After dichotomization, the Markov transition model with random intercepts was applied. Overall analysis results depict a consistent image in supporting the favorable efficacy of the contingency management.

Our previous endeavors in methodology development for incomplete longitudinal data analysis mainly focused on software development and MCMC-based Bayesian computations. Simulation studies and practical applications for full-likelihood models have proved their acceptable performance; see Yang and Shoptaw (2005), Yang et al (2005), Li et al. (2005), and Yang & Li (2005). Currently, we are conducting more simulation studies to evaluate and compare the selection, pattern-mixture, and shared-parameter models. We developed the software package MPI 2.0 to implement all the three modeling functions within the framework of one or two stage MPI; see www.Bayessoft.com/MPI. This package also provides tools for visual data exploration and formal assessment on missing-data assumptions; see an example from Yang and Shoptaw (2005). The first two chapters of this technical report cover a wide range of topics related to this issue. Currently, we are implementing more functions to this package so that most types of repeated measures with distribution from the exponential family can be modeled.

When describing the hybrid Gibbs sampler for model fitting and imputation, only the general structure and ideas were presented. For details related to the theoretical basis and technical implementation, refer to Yang and Li (2005) and Li et al. (2005). Results of simulation studies for selection models and random-intercept Markov transition models are also presented in these two articles.

In preparing this report and carrying out the research on methodology and software development, many people contributed to our work. We especially thank the research team at Bayessoft, Inc., mainly including Xiaowei Yang (the P.I.), Steven Shoptaw (the Co-P.I.), Thomas Belin, Robert Jennrich, John Boscardin, Junhui Li, Gang Liu, Ming-Jen Wang, and Qing J. Zhang. We also appreciate the kind support from National Institute on Drug Abuse and continuous encouragements and advices from Dr. Larry Seitz (the project officer).
REFERENCE


Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65:1, 275-297


