Chapter 1
Statistical Models and Methods for Incomplete Data in Randomized Clinical Trials

Michael A. McIsaac and Richard J. Cook

Abstract In this chapter we discuss several models by which missing data can arise in clinical trials. The likelihood function is used as a basis for discussing different missing data mechanisms for incomplete responses in short-term and longitudinal studies, as well as for missing covariates. We critically discuss common ad hoc strategies for dealing with incomplete data, such as complete-case analyses and naive methods of imputation, and we review more broadly appropriate approaches for dealing with incomplete data in terms of asymptotic and empirical frequency properties. These methods include the EM algorithm, multiple imputation, and inverse probability weighted estimating equations. Simulation studies are reported which demonstrate how to implement these procedures and examine performance empirically.

1.1 Introduction

In well-conducted randomized clinical trials, randomization eliminates the possible effect of confounding variables in the assessment of treatment effects. That is, when the assignment of the treatment to patients is carried out by random allocation, different treatment groups will have similar distributions of demographic and clinical features, so any differences seen in the distribution of responses between the treatment groups are attributable to the different treatments they receive. There are a number of other rationale put forward for use of randomization in health research [40], but it is the elimination of the effect of confounding variables and facilitation of causal inference that has had the most profound impact in advancing scientific understanding.

Following recruitment and randomization, however, participants in clinical trials often withdraw before completion of follow-up, leading to incomplete outcome

M.A. McIsaac (✉)
Department of Public Health Sciences, Queen's University, Kingston, ON, Canada
e-mail: mcisaacm@queensu.ca

R.J. Cook
Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada
e-mail: rjcook@uwaterloo.ca

K. van Montfort et al. (eds.), *Developments in Statistical Evaluation of Clinical Trials*, DOI 10.1007/978-3-642-55345-5_1

data. Incomplete data can of course arise for a variety of reasons; many illustrative examples can be seen in the second chapter of Molenberghs and Kenward [26]. Depending on the reasons for withdrawal, the individuals who remain in the study may no longer form groups with similar distributions of the demographic and clinical features, which compromises the validity of causal inferences. The purpose of this article is to discuss models and mechanisms by which incomplete data can arise in clinical trials, the consequences missing data can have on the interpretation of study results, and methods which can be employed to minimize the effect of these consequences. A clear understanding of the practical and statistical issues involved with incomplete response data will improve ability to critically appraise the clinical literature.

The remainder of this chapter is organized as follows. In Sect. 1.2 we discuss the problem of incomplete binary responses. We restrict attention to the case of a binary treatment indicator and a single binary confounding variable to simplify the discussion, calculations, and empirical studies, but we remark on practical issues with more complex settings at the end of this section. We discuss the case of incomplete longitudinal data in Sect. 1.3, and the problem of incomplete covariates in Sect. 1.4. Concluding remarks are made in Sect. 1.5.

1.2 Incomplete Binary Response Data

1.2.1 Models and Measures of Treatment Effect

Consider a balanced two-arm clinical trial in which patients are randomized to receive either an experimental treatment or standard care. Let $X = 1$ indicate that a patient was allocated to receive experimental therapy and $X = 0$ otherwise, where $P(X = 1) = 0.5$. Suppose the outcome of interest is whether the patient had a successful response; we let $Y = 1$ if this is the case and $Y = 0$ otherwise. We illustrate the problem of dependently missing data by considering a situation with a single additional binary variable V, where $V = 1$ indicates the presence of a particular feature and $V = 0$ otherwise; $P(V = 1) = p$. Suppose that the variable V is an effect modifier [33] so that the treatment has a different effect for individuals with and without the feature. This may be represented by the logistic model

$$P(Y = 1|X, V; \gamma) = \text{expit}(\gamma_0 + \gamma_1 X + \gamma_2 V + \gamma_3 XV) \,, \tag{1.1}$$

where $\gamma = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)'$. In most situations there will be sub-populations between which there is variation in the event rate and the effect of treatment; (1.1) is the simplest model which accommodates this phenomenon.

While (1.1) may reflect reality, in clinical trials we typically aim to assess treatment effects based on marginal models (i.e. models that do not condition on prognostic variables such as V); indeed provided X is independent of V, the causal

effect of treatment is typically defined in terms of such a model. Thus the logistic model used for treatment comparisons is formulated as

$$P(Y = 1|X;\beta) = \text{expit}(\beta_0 + \beta_1 X) \ , \tag{1.2}$$

where $\beta = (\beta_0, \beta_1)'$. Of course,

$$P(Y = 1|X;\beta) = E_V \left[P(Y = 1|X, V;\gamma); p\right] \ , \tag{1.3}$$

since V is independent of X due to randomization, and so it is possible to obtain the functional form of β in terms of $(\gamma', p)'$.

The resulting response rates in the control and treatment arms are $p_C = P(Y = 1|X = 0) = \text{expit}(\beta_0)$ and $p_T = P(Y = 1|X = 1) = \text{expit}(\beta_0+\beta_1)$, respectively. Some common measures of treatment effect include the absolute difference AD $= p_T - p_C$, the number needed to treat NNT $= (p_T - p_C)^{-1}$, the relative risk RR $= p_T/p_C$, and the odds ratio OR $= [p_T/(1-p_T)]/[p_C/(1-p_C)]$ [16,22]. When the experimental treatment has a higher response rate, the AD and NNT measures are positive and the RR and OR are larger than one.

Let $I(A)$ be an indicator function such that $I(A) = 1$ if A is true and $I(A) = 0$ otherwise. If response data are incomplete, in order to thoroughly discuss modeling issues it is necessary to introduce a new random variable $R = I(Y \text{ observed})$, so $R = 1$ if Y is observed and $R = 0$ otherwise. The biases that result from incomplete data arise if there is an association between the response (Y) and whether we observe it or not (R). There are a variety of ways of introducing an association between Y and R including through bivariate binary models [6] and shared random effect models [1]. Here we consider the setting in which both Y and R are associated with the covariates X and V. When V is unknown, an association between Y and R exists because of the omission of V from the analysis. We adopt this framework because when V is known, there are a variety of approaches to incorporating information about V into the analyses to mitigate problems, as we discuss in the following sections.

Suppose that the missing data model is

$$P(R = 1|X, V;\alpha) = \text{expit}(\alpha_0 + \alpha_1 X + \alpha_2 V + \alpha_3 XV) \ , \tag{1.4}$$

where $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)'$. This model accommodates a different dependence on V in the two treatment arms. We assume in this idealized setting that $R \perp Y|X, V$. Since $X \perp V$ by randomization, the marginal proportion of missing data is

$$\begin{aligned} p_R = P(R = 1;\alpha, p) &= E_X \left\{E_V \left[P(R = 1|X, V)\right]\right\} \\ &= \sum_{x=0}^{1}\sum_{v=0}^{1} P(R = 1|X = x, V = v;\alpha)\ P(V = v;p)\ P(X = x) \ , \end{aligned}$$

where $P(V = v;p) = p^v(1-p)^{1-v}$, and $P(X = x) = 1/2$ if randomization is balanced. The joint probability mass function for $Y, R|X$ is

$$P(Y, R|X;\theta) = E_V\left[P(Y|X,V;\gamma)\ P(R|X,V;\alpha)\right] = \sum_{v=0}^{1} P(Y|X,V=v;\gamma)\ P(R|X,V=v;\alpha)P(V=v;p)\ , \tag{1.5}$$

where $\theta = (\alpha', \gamma', p)'$. From (1.5) we can derive the conditional odds ratio for the association between Y and R given X as

$$OR_{Y,R|X} = \frac{P(Y=1,R=1|X;\theta)}{P(Y=1,R=0|X;\theta)} \Big/ \frac{P(Y=0,R=1|X;\theta)}{P(Y=0,R=0|X;\theta)}\ ,$$

and we can calculate the conditional probability

$$P(Y|X,R;\theta) = \frac{P(Y,R|X;\theta)}{P(R|X;\theta)} = \frac{P(Y,R|X;\theta)}{\sum_{y=0}^{1} P(Y=y,R|X;\theta)}\ . \tag{1.6}$$

So, thus far we have defined a simple model for $Y|X,V$ and $R|X,V$ under the assumption that Y and R are conditionally independent given (X,V). When we condition on X but not V, the response Y and the missing data indicator R are associated (i.e. dependent). We have mentioned that this setting was problematic, but here we will explore why this is the case.

1.2.2 Parameter Estimation with Incomplete Response Data

1.2.2.1 Complete-Case Analyses

Complete-Case Analyses when Covariate V Is Unknown

The likelihood function is perhaps the most fruitful starting point when considering inference based on parametric models [39]. When response data may be incomplete, the availability of the response of interest is stochastic, and hence the observed data likelihood is

$$L \propto P(Y,R=1|X)^R\ P(R=0|X)^{1-R}\ .$$

Noting that $P(Y,R=1|X) = P(Y|R=1,X)P(R=1|X)$, this may be re-expressed as $L_{Y|R=1,X} \cdot L_{R|X}$ where

$$L_{Y|R=1,X} = \left[P(Y=1|R=1,X)^Y\ P(Y=0|R=1,X)^{1-Y}\right]^R \tag{1.7}$$

is obtained from $P(Y|R=1,X)^R$ by considering the two possible realizations of Y, and

$$L_{R|X} = P(R=1|X)^R \, P(R=0|X)^{1-R} \, . \tag{1.8}$$

When responses are not available from all individuals in a sample, it is tempting to restrict attention to individuals with complete data and base analyses on this subset. This restriction, however, implicitly conditions on $R=1$ so that a complete-case maximum likelihood analysis actually maximizes the partial likelihood (1.7). It appears that (1.8) does not contain information about the parameters we are interested in because it relates to the missing data process alone. Note however that while (1.7) is indexed by θ, the quantities estimated by standard analyses based on available data (i.e. the sub-sample of individuals with $R=1$) are

$$\beta_0^\dagger = \text{logit}\, P(Y=1|X=0,R=1;\theta)$$

and

$$\beta_1^\dagger = \text{logit}\, P(Y=1|X=1,R=1;\theta) - \beta_0^\dagger \, .$$

These parameters differ from β_0 and β_1 whenever $P(Y|X,R=1) \neq P(Y|X)$, which will occur here if $P(Y|X,V) \neq P(Y|X)$ and $P(R|X,V) \neq P(R|X)$. Using (1.6), we can compute the naive measures of treatment effect which are actually being estimated from complete-case analyses: $\text{AD}^\dagger = P(Y=1|X=1,R=1) - P(Y=1|X=0,R=1)$, $\text{NNT}^\dagger = 1/\text{AD}^\dagger$, $\text{RR}^\dagger = P(Y=1|X=1,R=1)/P(Y=1|X=0,R=1)$, and $\text{OR}^\dagger = [P(Y=1|X=1,R=1)/P(Y=0|X=1,R=1)]/[P(Y=1|X=0,R=1)/P(Y=0|X=0,R=1)]$.

To explore this more fully, we consider here some specific parameter configurations. Let $P(X=1)=0.5$ and $P(V=1)=0.5$. In the response model (1.1), we let $\gamma_2=0$ and $\gamma_3=\log 2$ so the odds ratio characterizing the treatment effect is twice as big for those with $V=1$ compared to those with $V=0$. We set $\beta_1=\log 1.5$ in (1.2), so the marginal odds ratio of the treatment effect is 1.5, and we solve for γ_0 and γ_1 so that $P(Y=1|X=0)=\text{expit}(\beta_0)=0.5$ (i.e. the probability of response is 0.5 in the control arm). The marginal relative risk is therefore 1.2. In the missing data model (1.4) we set $\alpha_1=\alpha_2=0$ and for each α_3 we solve for α_0 so that $P(R=1)=0.5$.

Figure 1.1 displays a plot of $RR^\dagger$ and $OR^\dagger$, the limiting values of complete-case estimators of RR and OR, as a function of α_3. When $\alpha_3=0$, the probability of the response being missing is the same for all individuals regardless of their covariates (data are *missing completely at random*, in the terminology of Little and Rubin [20]), so $P(R|X,V)=P(R|X)=P(R)$. In this case, $RR^\dagger=RR=1.2$ and $OR^\dagger=OR=1.5$. When $\alpha_3<0$, complete-case estimators of these effect measures will be too small and hence correspond to a understatement of the effect

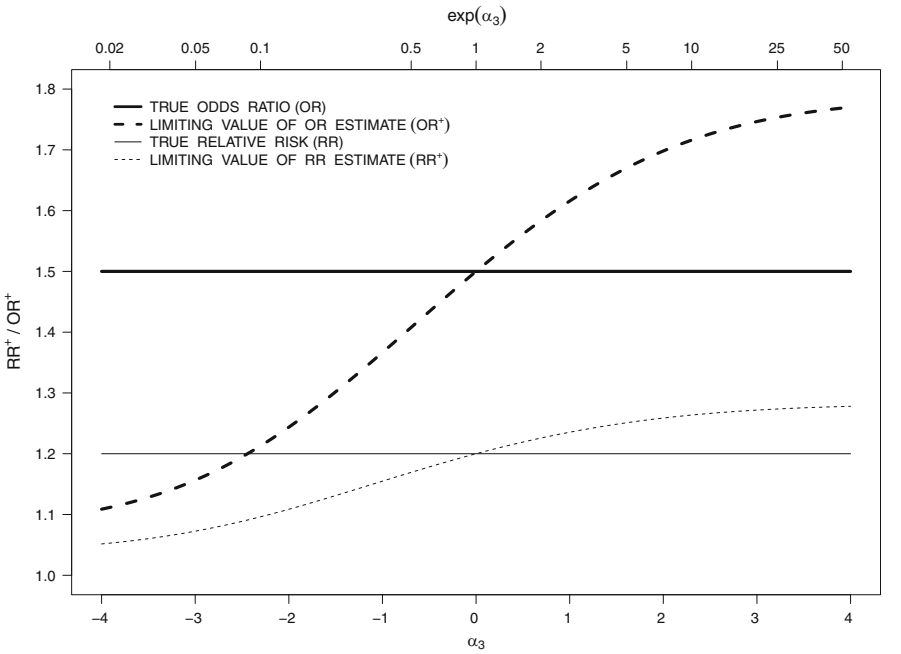

Fig. 1.1 Limiting values of naive complete-case estimators of the relative risk ($RR^{\dagger}$) and odds ratio ($OR^{\dagger}$) as a function of α_3

of treatment. Conversely, when $\alpha_3 > 0$, the inferences regarding the benefit of treatment are anti-conservative.

Complete-Case Analyses when Covariate V Is Known

If we are able to identify the variable V which renders Y and R conditionally independent (i.e., $Y \perp R|X, V$), another option is to write the observed data likelihood based on the conditional model as

$$L \propto P(Y, R = 1|X, V)^R \left[P(R = 0|X, V)\right]^{1-R} .$$

Since $P(Y, R = 1|X, V) = P(Y|X, V)P(R = 1|Y, X, V)$ and $P(R = 1|Y, X, V) = P(R = 1|X, V)$ this can in turn be written as $L_{Y|X,V} \cdot L_{R|X,V}$ where $L_{Y|X,V} \propto P(Y|X, V)$ and $L_{R|X,V} \propto P(R|X, V)$. In practice one would naturally restrict attention to the partial likelihood $L_{Y|X,V}$, since we are not typically interested in modeling the missing data process unless it is necessary. As seen above, a complete-case analysis with restriction to individuals with $R = 1$ yields inconsistent estimators of β when we just condition on X, however when we

condition on V as well, a complete-case analysis gives consistent estimators for γ. Identification of variables like V which are prognostic for Y and associated with the missing data process is therefore key to ensure consistent estimation of parameters. It is not sufficient for these variables to be associated with the response alone or the missing data status alone since in either case such variables cannot render Y and R conditionally independent.

While conditioning on a suitable V seems to have solved our problem, the catch is that we did not want to condition on V in our assessment of the treatment effect – we are estimating γ instead of β, so we are estimating the wrong thing! We do have the option of modeling $V|X$, which amounts to modeling the marginal distribution of V since X was determined by randomization, and given an estimate of p as $\hat{p}$, we can compute a crude estimate by solving for β in

$$\tilde{P}(Y = 1|X; \tilde{\beta}) = \sum_{v=0}^{1} P(Y = 1|X, V = v; \hat{\gamma})\ \hat{p}^v(1-\hat{p})^{1-v}\ .$$

Due to the so-called curse of dimensionality, this process is considerably more challenging and undesirable when V is high dimensional (i.e. a vector) [30]. A very convenient and more direct approach to estimating β is obtained using inverse probability weights as we describe in the next sub-section.

1.2.2.2 Use of Inverse Probability Weights

Suppose we have a sample of n independent subjects giving data $\{(Y_i, X_i, V_i), i = 1, 2, \ldots, n\}$. The score function for the logistic regression model in (1.2) resulting from (1.7) can be written as

$$S(\beta) = \sum_{i=1}^{n} R_i\ \ (Y_i - E(Y_i|X_i; \beta))\ \begin{pmatrix} 1 \\ X_i \end{pmatrix}\ .$$

With complete data (i.e. if $P(R_i = 1) = 1, i = 1, 2, \ldots, n$) this has expectation zero and hence yields a consistent estimator for β [23]. With incomplete data however,

$$\begin{aligned} E[S(\beta)] &= E_X\left\{E_{Y|X}\left\{E_{R|Y,X}\left[S(\beta)\right]\right\}\right\} \\ &= \sum_{i=1}^{n} E_X\left\{E_{Y|X}\left[P(R_i = 1|Y_i, X_i)\ (Y_i - E(Y_i|X_i; \beta))\begin{pmatrix} 1 \\ X_i \end{pmatrix}\right]\right\}\ , \end{aligned}$$

which does not in general equal zero. If the probability of a response being missing depends on Y given X, then inconsistent estimators are obtained for β; the corresponding limiting values are the $\beta^\dagger$ given in the previous section.

Now again suppose we are able to identify V as a covariate which renders $Y \perp R|X, V$. In this case we can employ the model for $P(R = 1|Y, X, V) = P(R = 1|X, V; \alpha)$ in an *inverse probability weighted* estimating function defined as

$$U(\beta) = \sum_{i=1}^{n} \frac{R_i}{P(R_i = 1|X_i, V_i; \alpha)} (Y_i - E(Y_i|X_i; \beta)) \begin{pmatrix} 1 \\ X_i \end{pmatrix} \tag{1.9}$$

[32]. Taking the expectation of (1.9) as before yields

$$\begin{aligned} E[U(\beta)] &= \sum_{i=1}^{n} E_{X,V} \left\{ E_{Y|X,V} \left[E_{R|Y,X,V} \left(\frac{R_i}{P(R_i = 1|X_i, V_i)} (Y_i - E(Y_i|X_i; \beta)) \begin{pmatrix} 1 \\ X_i \end{pmatrix} \right) \right] \right\} \\ &= \sum_{i=1}^{n} E_{X,V} \left\{ E_{Y|X,V} \left[(Y_i - E(Y_i|X_i; \beta)) \begin{pmatrix} 1 \\ X_i \end{pmatrix} \right] \right\} \\ &= \sum_{i=1}^{n} E_X \left\{ E_{V|X} \left\{ (E(Y_i|X_i, V_i) - E(Y_i|X_i; \beta)) \begin{pmatrix} 1 \\ X_i \end{pmatrix} \right\} \right\} \\ &= \sum_{i=1}^{n} E_X \left[(E(Y_i|X_i; \beta) - E(Y_i|X_i; \beta)) \begin{pmatrix} 1 \\ X_i \end{pmatrix} \right] = 0 \end{aligned} \tag{1.10}$$

and so a consistent estimator of β is obtained from (1.9) [11].

Note that in practice the parameters in the model $P(R|X, V; \alpha)$ must be estimated and this can easily be carried out via logistic regression since R is a binary variable. Naive standard errors which do not recognize that the weights have been estimated can lead to invalid tests (with incorrect type I error rates) and invalid confidence intervals (with coverage rates not compatible with the nominal level). Large sample theory for correct variance estimation is beyond the scope of this note, but see Robins et al. [32] for general results or Chen and Cook [3] for simpler results corresponding to the present formulation.

1.2.2.3 Multiple Imputation

Multiple imputation is, in its simplest implementation, a simulation-based approach to creating complete data from an incomplete dataset. Again suppose that we have identified a covariate V which renders $Y \perp R|X, V$, and the model for $Y|X, V$ is given by (1.1). A multiple imputation approach involves fitting a model to $Y|X, V$ based on individuals with complete data, even though $Y|X$ is the model of interest. The fitted model would give a consistent maximum likelihood estimator $\hat{\gamma}$, along with the asymptotic covariance matrix for $\hat{\gamma}$, $\mathscr{I}^{-1}(\hat{\gamma})$, where $\mathscr{I}(\gamma)$ is the expected information matrix from an analysis based on (1.1). Since γ is not of interest, this fitted model is simply used to generate complete data which are then analyzed with

the model of interest. The particular steps in such analyses are described in the following paragraphs.

The approach has a Bayesian flavour in that after fitting $Y|X, V$ we sample from MVN$(\hat{\gamma}, \mathscr{I}^{-1}(\hat{\gamma}))$ to obtain another realization of the 4×1 parameter vector $\hat{\gamma}$ which we denote by $g^{(1)}$. If the response for any individual is missing, then we simulate the binary response as a Bernoulli variate with probability expit$(g_0^{(1)} + g_1^{(1)}X + g_2^{(1)}V + g_3^{(1)}XV)$ using the respective covariate values. This yields *the first imputed value* for each individual with missing data, and we label the realized response $y^{(1)}$. After each individual with incomplete data in the dataset has a response simulated based on $g^{(1)}$, a second sample is drawn from MVN$(\hat{\gamma}, \mathscr{I}^{-1}(\hat{\gamma}))$ and labelled $g^{(2)}$. Using this value, one samples a second value $Y^{(2)} \sim$ Bern(expit$(g_0^{(2)} + g_1^{(2)}X + g_2^{(2)}V + g_3^{(2)}XV)$) for each person with a missing response data. This procedure is repeated m times until we have m "complete" datasets. For each of the m "complete" datasets we then fit the model of interest given by (1.2).

Let $\hat{\beta}_1^{(r)}$ denote the estimate of β_1 from the rth imputed data set and $\omega^{(r)} = \widehat{\text{var}}(\hat{\beta}_1^{(r)})$ be the naive variance estimate ignoring the fact that some data had been imputed by simulation. The combined estimate of β_1 obtained by multiple imputation is simply the average, so $\bar{\hat{\beta}}_1 = \sum_{r=1}^m \hat{\beta}_1^{(r)}/m$ is the reported point estimate from multiple imputation. Let $\bar{\omega} = \sum_{r=1}^m \omega^{(r)}/m$ denote the average of the naive (within imputation) variance estimates, and let $\omega^* = (m-1)^{-1}\sum_{r=1}^m(\hat{\beta}_1^{(r)} - \bar{\hat{\beta}}_1)^2$ denote the variation between imputation samples. Rubin [36] argues that the asymptotic variance of $\bar{\hat{\beta}}_1$ is $\text{var}(\bar{\hat{\beta}}_1) = \bar{\omega} + (1 + m^{-1})\omega^*$ and

$$\frac{\bar{\hat{\beta}}_1 - \beta_1}{\sqrt{\text{var}\left(\bar{\hat{\beta}}_1\right)}} \sim t_{u_m}$$

approximately, where the degrees of freedom are given by

$$u_m = (m-1)\left[1 + \frac{m\bar{\omega}}{(1+m)\omega^*}\right]^2.$$

Wang and Robins [42] prove consistency and derive the large sample properties of estimators arising from multiple imputation under correct model specification. More refinements to the estimated degrees of freedom have since been made [2] and are implemented in SAS. We will not get into these issues here, but remark simply that one appeal of multiple imputation is the ability to make use of auxiliary variables such as V when constructing the imputation model. In the context of longitudinal data with missing at random processes (see Sect. 1.3), this can be achieved by adopting a joint model for the responses over time (e.g., a mixed model) and, while the primary analysis is to be based only on a final response, intermediate values

can ensure a more suitable imputation process which may translate to more precise estimates of treatment effects and more powerful tests.

1.2.3 An Illustrative Simulation Study

Here we report on a simple simulation study to illustrate these methods. We let $p_C = 0.5$, $P(V = 1) = p = 0.5$, $\beta_1 = \log 1.5$, $\gamma_2 = \log 0.5$ and $\gamma_3 = \log 2$. These specifications can be used to obtain values for γ_0 and γ_1. Note that the true odds ratio $\exp(\beta_1)$, which would be consistently estimated in the absence of missing data, is 1.5 in this formulation ($\beta_1 \approx 0.4055$). We then specify the missing data model as $\alpha_1 = 0$, $\alpha_2 = \log 2$, $\alpha_3 = \log 4$, and ensure that $P(R = 1) = p_R = 0.5$, so 50 % of subjects will have incomplete response data and there is a differential degree of association between Y and R in the control and treatment arms. The limiting value of a naive estimate of β_1 is 0.4831 based on the earlier calculations, giving an asymptotic bias of approximately 0.0777.

Two thousand datasets of $n = 500$ individuals were simulated and the following analyses were carried out: (i) a complete-case likelihood analysis using (1.7), (ii) an inverse weighted analysis using (1.9) with weights known, (iii) an inverse weighted analysis with weights estimated via logistic regression, and (iv) multiple imputation with $m = 20$ and the imputation model based on $Y|X, V$. In all cases the response model was simply based on $Y|X$. The empirical biases, empirical standard errors (ESE), average asymptotic standard errors (ASE), and empirical coverage of nominal 95 % confidence intervals (ECP) are reported in Table 1.1.

The empirical biases of the complete-case analyses (expected since $\gamma_3 \neq 0$ and $\alpha_3 \neq 0$) are apparent, and this leads to empirical coverage probabilities less than the nominal 95 % level. The bias from the inverse weighted analyses

Table 1.1 Simulation results of naive and adjusted analyses using inverse weighting (known and estimated weights) and multiple imputation; $P(X = 1) = 0.5$; $P(V = 1) = 0.5$; $p_C = 0.5$; $\beta_0 = 0$, $\beta_1 = \log 1.5$, $\gamma_0 = 0.347$, $\gamma_1 = 0.059$, $\gamma_2 = \log 0.5$,$\gamma_3 = \log 2$,$p_R = 0.5$; $\alpha_0 = -0.654$, $\alpha_1 = 0$, $\alpha_2 = \log 2$, $\alpha_3 = \log 4$, Number of subjects = 500; Number of simulations = 2,000

Method of analysis	Parameter	Bias	ESE	ASE	ECP
Complete-case analysis	β_0	−0.072	0.201	0.196	93.3
	β_1	0.076	0.268	0.260	93.1
Weighted analysis	β_0	−0.005	0.204	0.199	95.1
(Known weights)	β_1	0.009	0.278	0.274	94.1
Weighted analysis	β_0	−0.004	0.203	0.200	95.2
(Estimated weights)	β_1	0.008	0.279	0.275	94.3
Multiple imputation[a]	β_0	−0.004	0.203	0.195	94.2
($m = 20$)	β_1	−0.004	0.281	0.277	94.2

[a] m indicates the number of complete pseudo-datasets created for multiple imputation

with known and estimated weights are negligible and the empirical coverage probabilities are compatible with the 95 % level. The biases are similarly small for the estimators based on multiple imputation and the empirical coverage probabilities are compatible with the 95 % level for these as well. Also noteworthy is the similarity in the standard errors of the estimates based on inverse weighting and multiple imputation.

1.2.4 Further Remarks

In many clinical settings there are a number of ad hoc alternative approaches for dealing with missing response data. In dermatology trials, for example, it is common to use so-called *non-responder* imputation [12, 28]. If, as we have described here, the response $Y = 1$ indicates a successful response to treatment (e.g. alleviation of symptoms), then in non-responder imputation (NRI), individuals who do not provide a response are assigned a value $Y = 0$ (i.e. they did not remain in the trial and report an alleviation of symptoms). The rationale for this crude form of imputation may arise from the notion that anything other than completing the course of treatment and exhibiting a good clinical response is undesirable and hence should be treated as a failure. An intuitively appealing aspect of this form of imputation is that all patients randomized are utilized in the analysis. However with NRI, a naive estimator of the probability of a successful response given X is, in fact, consistent for the joint probability $P(Y = 1, R = 1|X)$; this reflects that individuals must both provide a response and the response must be successful. The validity of estimates achieved through this method depends, therefore, on the process giving rise to the missing data. If $R \perp (Y, X)$, estimates of response rates within treatment arms (and therefore also estimates of AD) are conservative in that they are down-weighted by the probability of a response being observed (in fact, we are consistently estimating $P(Y = 1|X) \cdot P(R = 1)$). When data are not missing completely at random, NRI analyses will not yield consistent estimates of RR, OR, or AD. Depending on the mechanism giving rise to the missing data (which is generally unknown), NRI analyses can lead to conservative (too small) or anti-conservative (too large) estimates of treatment effect [25]. Despite this, NRI is commonly assumed to be a conservative method of analysis [37].

When responses are continuous, the calculations discussed in previous sections can be carried out following similar principles; to make this clear we wrote the expressions in a general form using expectations and explicit probability statements in key places. With continuous responses, however, another common crude method of imputation is often used called *mean value* imputation. In this case the average value of the response (perhaps for that particular treatment arm, or overall) is assigned to individuals with missing responses. This strategy can also lead to conservative or anti-conservative estimates of treatment effect depending on the particular setting, and naive standard errors will not typically reflect the effect of imputation.

The discussion of multiple imputation given earlier is often referred to as *parametric* multiple imputation since it relies on the explicit specification of a parametric model to simulate the imputed data for each data set. Other versions of multiple imputation are often adopted which employ implicit models to exploit the data observed in the sample [15, 21]. *Nonparametric* multiple imputation involves finding a set of completely observed individuals who are "similar" to an individual with a missing response (with respect to key attributes or a summary measure) and randomly selecting the responses from this set of similar individuals [29, 38]. This sampling is done with replacement to make up multiple complete datasets. Here judgement is not required to specify a probability model for imputation of the response, but rather to identify the set of "similar" individuals for each individual with a missing response [36]. Matching, stratification or use of propensity scores are useful for this goal, and several procedures are available in common statistical packages to facilitate this.

1.3 Incomplete Longitudinal Data

1.3.1 Notation and Terminology

Consider a longitudinal study in which the plan is to assess each of n individuals over K distinct assessment times. Let $Y_i = (Y_{i1}, \ldots, Y_{iK})'$ denote the random variable corresponding to the response vector for individual i over the K assessments. Suppose that every individual under study has measurements taken on p baseline covariates so that subject i has baseline covariate vector $X_i = (X_{i1}, \ldots, X_{ip})'$. We assume X_i is completely observed, and let $P(Y_i|X_i)$ denote the probability model of interest.

We restrict attention to incomplete longitudinal data due to drop-out, and suppose that the last time an observation for individual i occurred was at time K_i; this is a random variable and we let k_i denote its realization, as illustrated in Fig. 1.2. We can then partition the response vector as $Y_i = (\bar{Y}_i, Y_i^-)$, where $\bar{Y}_i = (Y_{i1}, \ldots, Y_{iK_i})'$ is observed and $Y_i^- = (Y_{i,K_i+1}, \ldots, Y_{iK})$ is missing. Let $R_i = (R_{i1}, \ldots, R_{iK})'$ be the corresponding vector of missing data indicators, where $R_{ik} = I(k \leq K_i)$, $k = 1, \ldots, K$. We can therefore equivalently think of R_i as a random vector or K_i as a random variable. Little and Rubin [20] and Rubin [35] define three classes of missing data mechanisms for this context.

Data are said to be *missing completely at random (MCAR)* if missingness (failing to observe a value) does not depend on any observed or unobserved measurements, i.e. $P(R_i|Y_i, X_i) = P(R_i)$. Data are said to be *missing at random (MAR)* if, conditional on the observed data, missingness does not depend on the data that are unobserved; that is, $P(R_i|Y_i, X_i) = P(R_i|\bar{Y}_i, X_i)$. Data are said to be *not missing at random* or, equivalently, *missing not at random (MNAR)* if missingness

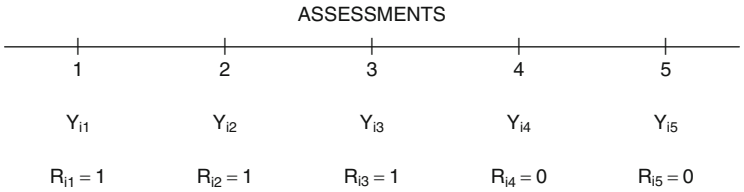

Fig. 1.2 Schematic of schedule of assessments in longitudinal study with $K = 5$ for an individual with $k_i = 3$

depends on the value of the realized (but unobserved) response, i.e. $P(R_i|Y_i, X_i)$ cannot be simplified. It is perhaps worth emphasizing that these terms must be used and interpreted in the context of the available information (or at least the information being used); MNAR mechanism can become a MAR mechanism in light of additional information used judiciously.

1.3.2 Likelihood-Based Methods of Estimation and Inference

As in the univariate case, the likelihood for incomplete longitudinal data is developed by specifying the joint distribution of response variable Y_i and the missing data indicators R_i (or equivalently K_i), given the covariates X_i. Two classes of models have been proposed based on alternative factorizations of the joint distribution of $(Y_i, R_i)|X_i$ [19]: one is based on *selection models* [20], the other is based on *pattern mixture models* [10, 18].

With selection models, the joint distribution of Y_i and R_i is factored as

$$P(R_i, Y_i|X_i; \beta, \alpha) = P(R_i|Y_i, X_i; \alpha)\ P(Y_i|X_i; \beta)\ , \tag{1.11}$$

where the distribution of R_i, $P(R_i|Y_i, X_i; \alpha)$, is indexed by a vector of parameters α and the distribution of Y_i, $P(Y_i|X_i; \beta)$, is indexed by a vector of β.

With pattern-mixture models, the factorization of the joint distribution is

$$P\ (R_i, Y_i|X_i; \beta, \alpha) = P\ (Y_i|X_i, R_i; \zeta)\ P\ (R_i|X_i; \theta)\ , \tag{1.12}$$

where in $P(Y_i|X_i, R_i; \zeta)$, the distribution of Y_i, is defined separately for each missing data configuration and indexed by parameters ζ, and the distribution of R_i, $P(R_i|X_i; \theta)$, is known up to parameters θ.

When we are concerned with the parameters of the marginal distribution of Y, averaged over the missing data patterns, it is in many senses more natural to use selection models, because people do not want to make inference conditional on the missing data indicators. In the following, we focus on selection models.

To describe the likelihood based approach we derive the joint density of the observed data $(\bar{Y}_i, R_i)$ by integrating out the missing data Y_i^- in the selection model of the joint distribution as

$$P(R_i, \bar{Y}_i | X_i; \alpha, \beta) = \int P(R_i | \bar{Y}_i, Y_i^-, X_i; \alpha)\ P(\bar{Y}_i, Y_i^- | X_i; \beta)\ dY_i^- .$$

Let $\bar{Y} = \{\bar{Y}_i, i = 1, 2, \ldots, n\}$ and $R = \{R_i, i = 1, 2, \ldots, n\}$ for a sample of n independent subjects. Then the observed-data joint likelihood for $(\alpha', \beta')'$ is

$$L(\alpha, \beta; \bar{Y}, R) = \prod_{i=1}^{n} \int P(R_i | \bar{Y}_i, Y_i^-, X_i; \alpha)\ P(\bar{Y}_i, Y_i^- | X_i; \beta)\ dY_i^- . \quad (1.13)$$

When the missing data mechanism is MAR, $P(R_i | \bar{Y}_i, Y_i^-, X_i) = P(R_i | \bar{Y}_i, X_i)$ and (1.13) becomes

$$\begin{aligned} L(\alpha, \beta; \bar{Y}, R) &= \prod_{i=1}^{n} \left\{ P(R_i | \bar{Y}_i, X_i; \alpha) \int P(\bar{Y}_i, Y_i^- | X_i; \beta)\ dY_i^- \right\} \quad (1.14) \\ &= \prod_{i=1}^{n} \left\{ P(R_i | \bar{Y}_i, X_i; \alpha)\ P(\bar{Y}_i | X_i; \beta) \right\} . \end{aligned}$$

If the parameters α and β are functionally independent, then likelihood inference for β from (1.14) is the same as a likelihood inference for β from the observed "partial" likelihood simply using the available data

$$L(\beta; \bar{Y}) = \prod_{i=1}^{n} P(\bar{Y}_i | X_i; \beta) . \quad (1.15)$$

Thus likelihood functions are unaffected by MAR mechanisms and this has contributed in part to the popularity of mixed effects models for the analysis of longitudinal data. If data are MNAR, then the simplification in (1.14) is not possible and we must use (1.13). This likelihood may lead to identifiability problems and so sensitivity analyses are often advocated for this case [31].

We remark that, as in the univariate case, one can sometimes identify an auxiliary covariate V_i which renders $R_i \perp Y_i^- | \bar{Y}_i, X_i, V_i$, so that inclusion of V_i in the analysis causes the missing data mechanism to be MAR. In this case, consider

$$\begin{aligned} P(R_i, \bar{Y} | X_i, V_i) &= \int P(R_i | \bar{Y}_i, Y_i^-, X_i, V_i)\ P(\bar{Y}_i, Y_i^- | X_i, V_i)\ dY_i^- \\ &= P(R_i | \bar{Y}_i, X_i, V_i)\ P(\bar{Y}_i | X_i, V_i) . \end{aligned}$$

This is only useful if we aim to estimate the effect of both X_i and V_i on the distribution of Y_i. Again, however, V_i may be useful for multiple imputation (as in Sect. 1.2.2.3) or for inverse weighting as we discuss in the next section.

1.3.3 Generalized Estimating Equations

Using standard notation for generalized linear models of binary data, we let $E(Y_{ik}|x_i) = P(Y_{ik} = 1|x_i) = \mu_{ik}$ and $\text{var}(Y_{ik}|x_i) = \mu_{ik}(1-\mu_{ik})$, $k = 1,\ldots,K$. Furthermore, we let $\Sigma_i(\beta,\rho) = \text{cov}(Y_i|x_i) = \mathbb{A}_i^{\frac{1}{2}} Q(\rho)\mathbb{A}_i^{\frac{1}{2}}$ where $\mathbb{A}_i = \text{diag}\{\mu_{ik}(1-\mu_{ik}), k = 1,\ldots,K\}$ and $Q(\rho)$ is a $K \times K$ working correlation matrix with (k,k') entry, $Q_{kk'}(\rho)$, parameterized in terms of a vector of association parameters ρ. A marginal generalized linear model is formed by letting $g(\mu_{ik}) = x'_{ik}\beta$ where $g(\cdot)$ is a known link function and $\beta = (\beta_0,\ldots,\beta_p)'$ is a $(p+1)\times 1$ vector of regression coefficients.

Generalized estimating equations for β take the form

$$U(\beta,\rho) = \sum_{i=1}^{n} U_i(\beta,\rho) = 0 \tag{1.16}$$

where $U_i(\beta,\rho) = G_i'(\beta)\Sigma_i^{-1}(\beta,\rho)(Y_i - \mu_i)$, with $\mu_i = (\mu_{i1},\ldots,\mu_{iK})'$ and $G_i(\beta) = \partial\mu_i(\beta)/\partial\beta'$ a $K \times (p+1)$ matrix of derivatives [17]. If $\hat{\beta}$ is the solution for fixed $\rho = \rho_o$, then asymptotically $\sqrt{n}(\hat{\beta}-\beta) \sim N(0, \text{var}(\sqrt{n}(\hat{\beta}-\beta)))$ with

$$\text{var}(\sqrt{n}(\hat{\beta}-\beta)) = [A^{-1}(\beta,\rho_o)][B(\beta,\rho_o)][A^{-1}(\beta,\rho_o)]', \tag{1.17}$$

where $A(\beta,\rho) = E(\partial U_i(\beta,\rho)/\partial\beta')$ and $B(\beta,\rho) = E(U_i(\beta,\rho)U_i'(\beta,\rho))$. When ρ is not specified, estimation of β is facilitated by iteratively replacing ρ with a $\sqrt{n}$-consistent moment-type estimate based on estimates of β at successive iterations of a scoring algorithm [17].

The functional form of $Q_{kk'}(\rho)$, $k \neq k', k, k' = 1,\ldots,K$, is typically unknown, but even if the correlation structure is misspecified, consistent estimators of β arise from solving (1.16), and (1.17) will still hold. However, misspecification of the correlation structure in (1.16) can lead to inefficient estimators of β and, in more extreme cases, problematic asymptotic properties arise for the solution [7]. In many cases, the working independence assumption can yield quite efficient estimators [41], so we set $Q_{kk'}(\rho) = \rho_o = 0$ for $k \neq k'$ in what follows. An estimate of (1.17) is obtained in this case by computing

$$\widehat{\text{var}}(\sqrt{n}(\hat{\beta}-\beta)) = [\hat{A}^{-1}(\hat{\beta},\rho_o)][\hat{B}(\hat{\beta},\rho_o)][\hat{A}^{-1}(\hat{\beta},\rho_o)]', \tag{1.18}$$

where

$$\hat{A}(\hat{\beta}, \rho_o) = -n^{-1} \sum_{i=1}^{n} G_i'(\hat{\beta}) \mathbb{A}_i^{-1}(\hat{\beta}, \rho_o) G_i(\hat{\beta}),$$

and

$$\hat{B}(\hat{\beta}, \rho_o) = n^{-1} \sum_{i=1}^{n} G_i'(\hat{\beta}) \mathbb{A}_i^{-1}(\hat{\beta}, \rho_o)(y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)' \mathbb{A}_i^{-1}(\hat{\beta}, \rho_o) G_i(\hat{\beta}) \ .$$

As in the univariate case, however, this estimating equation approach is not appropriate when data are incomplete and not missing completely at random.

Selection models provide a natural framework for characterizing factors which affect the risk of attrition in longitudinal studies. Let $R_{ik} = I(k \leq K_i)$ and $\bar{R}_{ik} = \{R_{i1}, \ldots, R_{ik}\}$, $k = 1, \ldots, K_i$. Selection models involve modeling the conditional probability of drop-out at each visit, which we denote here as $\eta_{ik} = P(R_{ik} = 0 | R_{i1} = \cdots = R_{i,k-1} = 1, y_i, x_i)$. As mentioned in Sect. 1.3.1, the nature of the relation between this conditional probability of drop-out, covariates, and (possibly missing) responses determines the impact that drop-outs have on inferences regarding the regression coefficients in the response model. We restrict attention here to settings in which data are MAR, with any covariate dependence based only on previously observed covariates or responses. In this case, η_{ik} may be a function of $\bar{Y}_i$ and X_i, but not of Y_i^-. Let $H_{ik}^y = \{y_{i1}, \ldots, y_{i,k-1}\}$ be the history of response Y up to time k. In practice, we typically let η_{ik} depend on H_{ik}^y and X_i.

Since R_{ik} is a binary variable it is convenient to formulate logistic regression models for the conditional probability of drop-out given by

$$\log(\eta_{ik}/(1 - \eta_{ik})) = w_{ik}' \alpha^{(k)} \ , \tag{1.19}$$

where $\alpha^{(k)} = (\alpha_0^{(k)}, \ldots, \alpha_{q_k}^{(k)})'$ is a $(q_k + 1) \times 1$ vector of regression coefficients characterizing the nature of the relationship between w_{ik} and η_{ik}, and w_{ik} is a covariate vector containing relevant observed information in H_{ik}^y and X_i.

The inverse-weighted estimating equations under the working independence assumption take the form

$$U(\beta, \alpha) = \sum_{i=1}^{n} U_i(\beta, \alpha) = 0 \tag{1.20}$$

where under cluster-specific weights as discussed by Fitzmaurice [9],

$$U_i(\beta, \alpha) = G_i'(\beta) \Sigma_i^{-1}(\beta) \Delta_i(\alpha)(Y_i - \mu_i) \ ,$$

$\Sigma_i(\beta) = \text{diag}\{\eta_{ik}(1 - \eta_{ik}), k = 1, \ldots, K_i\}$, $\Delta_i(\alpha) = I(K_i = k_i)/\pi_i(\alpha)$, and $\pi_i(\alpha) = P(K_i = k_i | \bar{Y}_i, x_i; \alpha)$. We often assume all subjects are available for the

first assessment, so $\pi_i(\alpha) = \eta_{i2}(\alpha)$ if $k_i = 1$, $\pi_i(\alpha) = (1 - \eta_{i2}(\alpha))\eta_{i3}(\alpha)$ if $k_i = 2$, $\pi_i(\alpha) = (1 - \eta_{i2}(\alpha))(1 - \eta_{i3}(\alpha))$ if $k_i = 3$, etc. In practice, an estimate of α can be obtained by fitting ordinary logistic regression models to the missing data indicators as appropriate. Inserting $\hat{\alpha}$ into (1.20) gives estimating equations which can be solved for β in the usual fashion [32].

1.3.4 Naive Methods of Imputation

The "last observation carried forward" (LOCF) imputation approach for dealing with missing values due to drop-outs operates as follows: if $k_i < K$, missing observations at visits $k = k_i + 1, \ldots, K$ are replaced with the value of the most recently observed response (i.e. y_{ik_i}). To distinguish the actual (possibly latent) responses from the pseudo-responses used under this imputation scheme, we use Y_i^* to denote the response vector under LOCF imputation. Therefore $Y_{ik}^* = Y_{ik}$ for $k \leq k_i$ and $Y_{ik}^* = Y_{ik_i}$ for $k > k_i$, $k = 1, 2, \ldots, K$. Assumptions made for the response Y_i are adopted for the pseudo-response Y_i^* since analyses are typically carried out under the assumption that they are in some sense equivalent. In fact, in most situations for which the assumptions regarding Y_i are true, they will not be true for Y_i^*, implying that the estimating equation (1.16) is misspecified for the pseudo response. The frequency properties of estimators of β based on Y_i^* have been investigated under a wide range of settings by several authors [5, 27] based on the theory of misspecified models [34, 43]. As with the other naive imputation approaches discussed earlier, LOCF leads to inconsistent estimators in a wide variety of settings and can result in either conservative or anti-conservative estimates of treatment effect.

1.4 Missing Covariates

1.4.1 Likelihood Analyses

Now consider a setting of a clinical trial in which the secondary analyses are directed at fitting a regression model which controls for a variable Z in addition to the treatment indicator; for the sake of simplicity we again suppose Z is a binary variable. One might simply specify a model with the main effects, but we consider a model of the form

$$P(Y = 1|X, Z; \lambda) = \text{expit}\,(\lambda_0 + \lambda_1 X + \lambda_2 Z + \lambda_3 XZ)\ . \tag{1.21}$$

This would be of interest if there are questions about whether the effect of treatment was significantly different in different subgroups defined by a binary covariate Z,

for example, in which case λ_3 is parameter of primary interest. Such questions arise frequently when the goal is to examine the robustness and generalizability of findings; in cancer trials, for example, the aim may be to investigate whether the effect of chemotherapy varies according to tumour type. Some centers may not collect complete histological data and in such circumstances covariate data on tumour type will be incomplete.

Let $C = I(Z \text{ observed})$ indicate whether the covariate value was recorded. The observed data likelihood can then be written as

$$L \propto P(Y, Z, C = 1|X)^C \; P(Y, C = 0|X)^{1-C} \; , \tag{1.22}$$

where we can marginalize over Z with $\sum_z P(Y, Z = z, C = 0|X)$ to obtain $P(Y, C = 0|X)$, the contribution from individuals for whom Z is unobserved.

As in the case of incomplete responses, the tendency is to focus on simple analyses such as those restricted to individuals with complete covariate data. In this case the adopted likelihood would be based on the response model with the implicit condition $C = 1$ and so is proportional to

$$\begin{aligned} P(Y|Z, X, C = 1) &= \frac{P(C = 1|Y, Z, X) \; P(Y|Z, X)}{\sum_y P(C = 1|Y = y, Z, X) \; P(Y = y|Z, X)} \\ &= \frac{P(C = 1|Y, Z, X)}{P(C = 1|Z, X)} \; P(Y|Z, X) \; . \end{aligned} \tag{1.23}$$

If $C \perp Y|Z, X$, then (1.23) reduces to $P(Y|Z, X)$ and a complete-case analysis will yield consistent estimators of λ, but otherwise inconsistent estimators are obtained; we show this by example in the simulation studies that follow. Note that with incomplete covariate data, missingness can depend on the potentially missing variable (Z) and a complete-case analysis remains valid because it involves conditioning on this covariate; this is in contrast to the setting of missing responses where the missing data must be modelled. However even when valid, this complete-case analysis ignores the information contained in the responses from individuals with incomplete data, and therefore may result in less than optimal efficiency.

1.4.2 An EM Algorithm

If one makes assumptions regarding the distribution of the incomplete covariate in likelihood analyses based on (1.22), one can exploit information from individuals with $C = 0$ and improve efficiency. To see this note that the second term in (1.22),

$$P(Y, C = 0|X) = \sum_{z=0}^{1} P(Y|Z = z, X) \; P(Z = z|X) \; P(C = 0|Y, Z = z, X) \; ,$$

is indexed by λ (as well as the parameters in $P(Z|X)$ and those of the missing data process). If $P(C|Y,Z,X) = P(C|Y,X)$ or $P(C|X)$, then the missing data process can be modelled using observed data (Y and X). If $P(C|Y,Z,X) = P(C|Z,X)$, then while this is a desirable missing data process for complete-case analysis (see (1.23)), in this setting there is a need to make uncheckable assumptions about the missing data process, since the dependence between C and Z given X cannot be modelled in general. Progress can be made here if an auxiliary variable V can be found which satisfies $C \perp Z|X,V,Y$ (see Sects. 1.4.3 and 1.4.4).

The assumptions that are needed to exploit information from individuals with $C = 0$ could include the fully specified conditional covariate distribution, or simply its parametric form. In the latter case, the EM algorithm offers a convenient method for estimation [8]. The complete data likelihood L_C corresponding to (1.22) is proportional to

$$[P(C|Y,Z,X)\ P(Y|Z,X)\ P(Z|X)]^C\ \ [P(C|Y,Z,X)\ P(Y|Z,X)\ P(Z|X)]^{1-C}\ .$$

We typically work with the "partial" complete data likelihood

$$L_C \propto [P(Y|Z,X)\ P(Z|X)]^C\ \ [P(Y|Z,X)\ P(Z|X)]^{1-C} \tag{1.24}$$

under the assumption that the information regarding λ in the missing-data model is negligible. Working with (1.24) then requires an expression for

$$P(Z|C=0,Y,X) = \frac{P(C=0|Y,Z,X)\ P(Y|Z,X)\ P(Z|X)}{\sum_z P(C=0|Y,Z=z,X)\ P(Y|Z=z,X)\ P(Z=z|X)} \tag{1.25}$$

for the expectation step of the EM algorithm, which if $C \perp Z|Y,X$ gives simply

$$\frac{P(Y|Z,X)\ P(Z|X)}{\sum_z P(Y|Z=z,X)\ P(Z=z|X)}\ . \tag{1.26}$$

It is clear from (1.26) that, provided $P(C|Y,Z,X) = P(C|Y,X)$, the partial complete data likelihood (1.24) can be used if assumptions are made regarding the distribution of $Z|X$. In fact, when treatment is randomly assigned, only the marginal distribution of Z is required since $Z \perp X$. However, if C depends on Z given Y and X, then there is an identifiability problem and (1.25) cannot be evaluated without strong assumptions regarding the missing data process.

1.4.3 Multiple Imputation with Missing Covariates

Suppose now that there exists a completely observed covariate V which renders $C \perp Z|Y,X,V$. Again for simplicity we assume V is binary with $P(V=1) = p$

and $P(V = 0) = 1 - p$. Multiple imputation can be carried out using a model for $P(Z|Y, X, V, C) = P(Z|Y, X, V)$ and because $Z \perp C|Y, X, V$, the model for $Z|Y, X, V$ can be fitted based on individuals with complete data. For illustration here, we adopt a simpler model whereby $P(Z|V, X, Y) = P(Z|V, X)$ which can be easily fitted using a saturated logistic regression model,

$$P(Z = 1|X, V) = \text{expit}(\delta_0 + \delta_1 X + \delta_2 V + \delta_3 XV) \,. \tag{1.27}$$

Suppose the missing data model is

$$P(C = 1|X, V) = \text{expit}(\alpha_0 + \alpha_1 X + \alpha_2 V + \alpha_3 XV) \,, \tag{1.28}$$

and the response is generated according to

$$P(Y = 1|X, Z, V) = \text{expit}(\lambda_0^* + \lambda_1^* X + \lambda_2^* Z + \lambda_3^* XZ + \lambda_4^* V) \,. \tag{1.29}$$

The response model of interest (1.21) can be recovered by noting that $P(Y = 1|X, Z) = E_{V|X,Z}[P(Y = 1|X, Z, V)]$.

The association between Y and C given X and Z is determined by the joint model

$$\begin{aligned} P(Y, C|X, Z) &= \sum_v P(Y|X, Z, V = v, C) P(C|X, Z, V = v) P(V = v|X, Z) \\ &= \sum_v P(Y|X, Z, V = v) P(C|X, V = v) P(V = v|Z) \,. \end{aligned}$$

If we simply fit the response model in (1.21), a complete-case analysis is generally invalid in this setting because $C \not\perp Y|X, Z$ due to the omission of the variable V in (1.21).

Following the same arguments as given earlier, for any given data set we may carry out multiple imputation of Z based on the model $P(Z|Y, X, V)$. If this model is fit and an estimate of δ is obtained, by standard large sample theory $\hat{\delta} \sim \text{MVN}(\delta, \mathscr{I}^{-1}(\hat{\delta}))$.

We proceed by letting $d^{(r)}$ denote the rth realization from $\text{MVN}(\hat{\delta}^*, \mathscr{I}^{-1}(\hat{\delta}^*))$, and using $d^{(r)}$ to generate values for all missing Z according to $P(Z|Y, X, V; d^{(r)})$. Then based on this "complete" data set, we fit $P(Y|Z, X; \lambda)$ to obtain $\hat{\lambda}^{(r)}$. This is repeated m times, and we let $\bar{\hat{\lambda}} = \sum_{r=1}^m \hat{\lambda}^{(r)}/m$ and compute the standard errors as described in Sect. 1.2.2.3.

1.4.4 Inverse Probability Weighted Estimating Functions

Inverse probability weighting can be used to obtain unbiased estimating functions for a complete-case analysis. If $P(C_i|Y_i, X_i, V_i, Z_i) = P(C_i|Y_i, X_i, V_i)$, then we can write the inverse weighted estimating function as

$$U(\beta) = \sum_{i=1}^{n} \frac{C_i}{P(C_i = 1|Y_i, X_i, V_i)} (Y_i - E(Y_i|X_i, Z_i; \lambda))\ W_i\ , \qquad (1.30)$$

where $W_i = (1, X_i, Z_i, X_i Z_i)'$, and this can be shown to have expectation zero. Since the model in the weight indicates a dependence on (Y_i, X_i, V_i) which are always observed, then it can be fit and a $\sqrt{n}$-consistent estimator of α in (1.28) inserted; a consistent estimator of λ will then be obtained by setting (1.30) equal to zero and solving for λ.

1.4.5 A Simulation Study

Here we report on a simulation study designed to demonstrate the performance of several methods of dealing with missing covariates. We consider the response model (1.21) with $\lambda_4^* = 0$ and $\log\ 4$ in (1.29) and find the parameters of the covariate distribution to ensure these parameter values were obtained. We set $\lambda_1 = 0$, $\lambda_2 = \log\ 1.5$, $\lambda_3 = \log\ 0.5$, $P(X = 1) = 0.5$, $P(V = 1) = 0.5$, and $P(Z = 1) = 0.25$ so $P(Y = 1) = 0.5$. We set $\delta_1 = 0$, $\delta_2 = 0$, $\delta_3 = \log\ 4$ in (1.27) to ensure that, as desired, $P(Z = 1) = 0.25$ based on (1.27). Finally, setting $\alpha_0 = -0.151$, $\alpha_1 = \log\ 0.8$, $\alpha_2 = \log\ 1.2$ and $\alpha_3 = \log\ 2$ in (1.28) yields $P(C = 1) = 0.5$; so for 50 % of subjects we would expect the covariate to be missing. We generated data for sample sizes of 500 and 2,000 individuals in 2,000 simulated datasets. The analyses conducted included a complete-case analysis, inverse probability weighted analyses with known and estimated weights, an EM algorithm for which the correct covariate distribution was assumed, and multiple imputation. The imputation model adopted was a saturated logistic regression model for Z given (Y, X, V), involving eight parameters: the intercept, three main effects, three two way interactions and a three way interaction. The empirical biases, empirical standard errors, average asymptotic standard errors, and empirical coverage probabilities are reported in Table 1.2 for sample sizes of 500 (left column) and 2,000 (right column). The top half of the table corresponds to the case where $C \perp Y|X, Z$; in the bottom half, $C \not\perp Y|X, Z$ but $C \perp Y|X, Z, V$ where V is the auxiliary covariate used for inverse weighting with $P(C|X, V)$, and multiple imputation via $P(Z|X, V)$.

The results where $Y \perp C|X, Z$ (top half) indicate all methods yield approximately unbiased estimates, close agreement between the empirical and average asymptotic standard errors, and empirical coverage that is compatible with the

Table 1.2 Simulation results of naive and adjusted analyses using inverse weighting (known and estimated weights), EM and multiple imputation; $P(X = 1) = 0.5$; $P(V = 1) = 0.5$; $P(Z = 1) = 0.25$, $\delta_1 = 0$, $\delta_2 = 0$, $\delta_3 = \log(4) = 1.3862$; $P(Y = 1) = 0.5$, $\lambda_1 = 0$, $\lambda_2 = \log(1.5) = 0.405$, $\lambda_3 = \log(0.5) = -0.693$; $P(C = 1) = 0.5$, $\alpha_0 = -0.151$, $\alpha_1 = \log(0.8) = -0.223$, $\alpha_2 = \log(1.2) = 0.182$, $\alpha_3 = \log(2) = 0.693$; Number of simulations $= 2,000$

Method	Parameter	Sample size: 500				Sample size: 2,000			
		Bias	ESE	ASE	ECP	Bias	ESE	ASE	ECP
		$Y \perp C \mid X, Z$ ($\lambda_4^* = 0$)							
Complete-case analysis	λ_0	0.001	0.203	0.202	95.2	0.001	0.101	0.100	95.1
	λ_1	−0.001	0.301	0.300	95.2	−0.005	0.152	0.149	94.3
	λ_2	0.017	0.531	0.505	95.2	0.015	0.248	0.244	94.9
	λ_3	−0.010	0.655	0.632	94.8	−0.015	0.315	0.307	94.7
Weighted analysis (Known weights)	λ_0	0.001	0.203	0.203	95.0	0.001	0.101	0.100	95.1
	λ_1	0.000	0.304	0.304	95.5	−0.006	0.152	0.151	94.7
	λ_2	0.017	0.532	0.506	95.5	0.015	0.248	0.244	95.0
	λ_3	−0.013	0.659	0.637	94.8	−0.015	0.317	0.310	94.6
Weighted analysis (Estimated weights)	λ_0	0.002	0.204	0.203	95.0	0.001	0.101	0.101	95.2
	λ_1	0.000	0.305	0.304	95.1	−0.006	0.152	0.151	94.8
	λ_2	0.018	0.534	0.507	95.5	0.015	0.248	0.244	95.2
	λ_3	−0.013	0.662	0.638	94.8	−0.016	0.317	0.310	94.5
EM	λ_0	−0.016	0.167	0.165	95.0	−0.016	0.081	0.082	94.9
	λ_1	0.010	0.240	0.238	94.6	0.011	0.118	0.118	95.0
	λ_2	−0.001	0.515	0.495	95.7	0.002	0.244	0.240	94.7
	λ_3	0.010	0.641	0.622	94.8	−0.001	0.310	0.304	94.5
Multiple imputation[a] ($m = 20$)	λ_0	0.007	0.157	0.158	95.4	0.002	0.075	0.077	95.6
	λ_1	−0.009	0.239	0.239	94.4	−0.003	0.116	0.116	95.0
	λ_2	−0.035	0.516	0.523	96.0	0.003	0.248	0.248	95.2
	λ_3	0.043	0.646	0.647	95.0	−0.010	0.315	0.311	94.5

		$Y \not\perp C \mid X, Z$ ($\lambda_4^* = \log(4)$)							
Complete-case analysis	λ_0	0.031	0.203	0.202	94.9	0.029	0.101	0.100	93.7
	λ_1	0.102	0.295	0.300	94.0	0.112	0.149	0.149	87.8
	λ_2	0.027	0.524	0.506	95.8	0.001	0.245	0.244	95.1
	λ_3	−0.084	0.647	0.633	95.8	−0.051	0.309	0.307	94.5
Weighted analysis (Known weights)	λ_0	0.000	0.203	0.203	95.1	−0.003	0.101	0.100	94.7
	λ_1	−0.004	0.300	0.304	95.6	0.005	0.150	0.151	95.5
	λ_2	0.027	0.525	0.507	95.7	0.001	0.245	0.245	95.1
	λ_3	−0.039	0.651	0.637	95.3	−0.008	0.311	0.309	95.0
Weighted analysis (Estimated weights)	λ_0	0.000	0.199	0.203	95.7	−0.003	0.099	0.101	95.0
	λ_1	−0.006	0.297	0.304	95.7	0.006	0.147	0.151	96.3
	λ_2	0.028	0.526	0.508	95.5	0.001	0.245	0.245	95.2
	λ_3	−0.037	0.651	0.638	95.2	−0.008	0.311	0.309	95.1
EM	λ_0	−0.009	0.166	0.165	94.8	−0.014	0.082	0.082	95.0
	λ_1	0.007	0.236	0.239	95.2	0.012	0.119	0.119	95.2
	λ_2	−0.007	0.512	0.496	95.4	−0.029	0.241	0.240	94.8
	λ_3	−0.005	0.638	0.622	95.3	0.026	0.305	0.303	95.0
Multiple imputation[a] ($m = 20$)	λ_0	0.010	0.155	0.163	95.9	0.000	0.076	0.077	95.8
	λ_1	−0.009	0.233	0.250	96.1	0.002	0.116	0.116	94.8
	λ_2	−0.024	0.500	0.544	96.8	−0.010	0.245	0.248	95.3
	λ_3	0.015	0.622	0.680	96.2	0.001	0.309	0.310	95.3

[a] m indicates the number of pseudo-complete datasets created for multiple imputation

nominal 95 % level. The efficiency gains realized by modeling the covariate distribution are apparent by comparing the standard errors from the complete-case analysis with those of the EM algorithm. The standard errors of the estimates from the EM and MI algorithms are in close agreement. For the bottom half of the table, the empirical biases from the complete-case analyses expected due to (1.23) are apparent. The weighted analyses yielded estimators with much smaller empirical biases and better performance with the larger sample size. Smaller biases and smaller standard errors are seen with the EM algorithm. The multiple imputation analyses yielded small empirical biases as well and their standard errors are in close agreement with those of the EM algorithm. The empirical coverage probabilities for all valid methods are compatible with the nominal 95 % level. Simulations and analyses were carried out in R version 2.14.0 and SAS 9.2 on the Sun Solaris 10 platform.

1.5 Discussion

Incomplete data can arise in a number of settings for a variety of different reasons. Key factors influencing the extent of the impact on standard analyses are the proportion of missing data, and as demonstrated in this chapter, the nature of the stochastic mechanism which causes the data to be incomplete. Even when analyses are valid, loss of efficiency and decreased power are always issues. When possible, the extent of missing data should always be minimized.

Likelihood methods which have been developed and applied to minimize the effect of incomplete data are often directed at retrieving information about parameters of interest and improving power, but these come at the cost of making modeling assumptions beyond those typically made in analyses with complete data. These additional model assumptions are explicit, for example, when a parametric multiple imputation approach is adopted for incomplete response data. When covariates are missing and the EM algorithm is applied, one must make assumptions regarding the covariate distribution, which is not customary in routine analyses. When inverse probability weights are used, a model for the missing data process must be specified, which again is not something that is routinely done in standard analyses. The specified models should be checked carefully since consistent estimators only result if these are correct.

Throughout this chapter we have emphasized simple models with binary data, primarily for transparency and so that explicit results would be easy to obtain. When responses are continuous, inverse probability weighting changes very little; this approach requires modeling the missing data indicator which remains binary. Multiple imputation can be carried out in this case based on a linear regression model. The methods for longitudinal data can be similarly adapted. When incompletely observed covariates are continuous or categorical, the necessary model assumptions for the EM algorithm or multiple imputation may become more involved and robustness of inferences becomes more of a concern. When multiple

covariates are missing, high-dimensional joint models for the covariates are required and these can be challenging to specify and check. These challenges, in part, are reasons for the appeal of inverse probability weighted analyses of individuals with complete data [24].

We have considered the cases of a missing response or a single missing covariate separately. Frequently both responses and covariates can be missing in a given dataset and hybrid methods can be employed [4].

We have emphasized the setting in which interest lies in a regression model for a marginal mean parameter. In some settings, association parameters (e.g. correlations or odds ratios) are viewed as of comparable importance. This occurs when scientific interest lies in the nature of the association structure, or if concerns lie in optimizing efficiency. In this case, regression models can be formulated for the association parameters and appropriate likelihood functions can be formed [13, 14]. Zhao and Prentice [45] describe how to do this using second order estimating equations. In the likelihood setting, the EM algorithm can be adopted and the idea of using inverse weighting for estimating association parameters can be adapted [44].

Acknowledgements This work was supported by a Post-Graduate Scholarship to Michael McIsaac from the Natural Sciences and Engineering Research Council (NSERC) of Canada and grants to Richard Cook from NSERC (Grant No. 101093) and the Canadian Institutes of Health Research (Grant No. 105099). Richard Cook is a Tier I Canada Research Chair in Statistical Methods for Health Research.

References

1. Albert, P.S., Follmann, D.: Shared-parameter models. In: Fitzmaurice, G., Davidian, M., Verbeke, G., Molenberghs, G. (eds.) *Longitudinal Data Analysis*, Chapter 18, 433–452. CRC Press, Boca Raton, FL. (2009)
2. Barnard, J., Rubin D.B.: Miscellanea. small-sample degrees of freedom with multiple imputation. *Biometrika* **86(4)**, 948–955 (1999)
3. Chen, B., Cook, R.J.: Strategies for bias reduction in estimation of marginal means with data missing at random. *Optimization and Data Analysis on Biomedical Informatics*. Ed: Panos Pardalos. American Mathematics Society (2011)
4. Chen, B., Yi, G.Y., Cook, R.J.: Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random. *Journal of the American Statistical Associaton* **105**, 336–353 (2010)
5. Cook, R.J., Zeng, L., Yi, G.Y.: Marginal analysis of incomplete longitudinal binary data: a cautionary note on LOCF imputation. *Biometrics* **60(3)**, 820–828 (2004)
6. Cox, D.R.: The analysis of multivariate binary data. *Applied Statistics* **21**, 113–120 (1972)
7. Crowder, M.: On the use of a working correlation matrix in using generalized linear models for repeated measures. *Biometrika* **82**, 407–410 (1995)
8. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society: Series B* **39**, 1–38 (1977)
9. Fitzmaurice, G.M., Molenberghs, G., Lipsitz, S.R.: Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society. Series B* (Methodological) **57(4)**, 691–704 (1995)

10. Glynn, R.J., Laird, N.M., Rubin, D.B.: Multiple imputation in mixture models for nonignorable nonresponse with followups. *Journal of the American Statistical Associaton* **88**, 984–993 (1993)
11. Godambe, V.P.: *Estimating Functions*. Oxford University Press, USA (1991)
12. Gordon, K.B., Langley, R.G., Leonardi, C., Toth, D., Menter, M.A., Kang, S., Hefferman, M., Miller, B., Hamlin, R., Lim, L., Zhong, J., Hoffman, R., Okun, M.M.: Clinical response to adalimumab treatment in patients with moderate to severe psoriasis: Double-blind, randomized controlled trial and open-label extension study. *Journal of the American Academy of Dermatology* **55**, 598–606 (2006)
13. Heagerty, P.J.: Marginalized transition models and likeliood inference for longitudinal categorical data. *Biometrics* **58**, 342–351 (2002)
14. Heagerty, P.J., Zeger, S.L.: Marginalized multilevel models and likelihood inference. *Statistical Science* **15**, 1–19 (2000)
15. Herzog, T., and Rubin, D.B.: Using multiple imputations to handle nonresponse in sample surveys. In: Madow, W.G., Olkin, I., Rubin, D.B. (eds.) *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliography*, 209–245. New York: Academic Press (1983)
16. Laupacis, A., Sackett, D.L., Roberts, R.S. : An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine* **318**, 1728–1733 (1988)
17. Liang, K.Y., Zeger, S. : Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22 (1986)
18. Little, R.J.A. : Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Associaton* **88**, 125–134 (1993)
19. Little, R.J.A. : Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Associaton* **90**, 1112–1121 (1995)
20. Little, R.J.A., Rubin, D.B.: *Statistical Analysis with Missing Data*. Wiley, New York (1987)
21. Little, R.J.A., Rubin, D.B.: *Statistical Analysis with Missing Data*, Second Edition. Wiley, New York (2002)
22. Matthews, D.E., Farewell, V.T.: *Using and Understanding Medical Statistics*, 3rd Revised Edition. Karger, Basel, Switzerland (1996)
23. McCullagh, P., Nelder, J.A. : *Generalized Linear Models*, Second Edition. Chapman & Hall/CRC., London, UK (1989)
24. McIsaac, M.A., Cook, R.J.: Response-Dependent Sampling with Clustered and Longitudinal Data. In: *ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers*, 157–181. New York: Springer (2013)
25. McIsaac, M.A., Cook, R.J., Poulin-Costello, M.: Incomplete data in randomized dermatology trials: Consequences and statistical methodology. *Dermatology* **226**(1), 19–27 (2013). DOI 10.1159/000346247
26. Molenberghs, G., Kenward M. : *Missing Data in Clinical Studies*. John Wiley & Sons Ltd, West Sussex, England, UK (2007)
27. Prakash, A., Risser, R. C., Mallinckrodt, C. H.: The impact of analytic method on interpretaion of outcomes in longitudinal clinical trials. *International Journal of Clinical Practice* **62**, 1147–1158 (2008)
28. Reich, K., Nestle, F.O., Papp, K., Ortonne, J.P., Evans, R., Guzzo, C., Dooley, L.T., Griffiths, C.E.M. for the EXPRESS Study Investigators : Infliximab induction and maintenance therapy for moderate-to-severe psoriasis: a phase III, multicentre, double-blind trial. *Lancet* **366**, 1367–1374 (2005)
29. Reilly, M., Pepe, M.: The relationship between hot-deck multiple imputation and weighted likelihood. *Statistics in Medicine* **16**, 5–19 (1997)
30. Robins, J.M., Ritov, Y.: Toward a curse of dimensionality approximate (CODA) asymptotic theory for semiparameric models. *Statistics in Medicine* **16**, 285–319 (1997)
31. Robins, J.M., Hernan, M.A., Brumback, B. Marginal structural models and causal inference in epidemiology. *Epidemiology* **11(5)**, 550–560 (2000)
32. Robins, J.M., Rotnitzky, A., Zhao, L.P.: Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90(429)**, 106–121 (1995)

33. Rothman, K.J., Greenland, S., eds: *Modern Epidemiology*, Second Edition. Lippincott Williams & Wilkins, Philadelphia (1998)
34. Rotnitzky, A., Wypij, D.: A note on the bias of estimators with missing data. *Biometrics* **50**, 1163–1170 (1994)
35. Rubin, D.B. : Inference and missing data. *Biometrika* **63**, 581–592 (1976)
36. Rubin, D.B. : *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York (1987)
37. Saurat, J.H., Stingl, G., Dubertret, L., Papp, K., Langley, R.G., Ortonne, J.P., Unnebrink, K., Kaul, M., Camez, A., for the CHAMPION Study Investigators: Efficacy and safety results from the randomized controlled comparative study of adalimumab vs. methotrexate vs. placebo in patients with psoriasis (CHAMPION). *British Journal of Dermatology* **158**, 558–566 (2007)
38. Schenker, N., Welsh, A.H.: Asymptotic results for multiple imputation. *The Annals of Statistics* **16(4)**, 1550–1566 (1988)
39. Sprott, D.A. : *Statistical Inference in Science*. Springer, New York (2000)
40. Sprott, D.A., Farewell, V.T.: Randomization in experimental science. *Statistical Papers* **34**, 89–94 (1993)
41. Sutradhar, B.C., Das, K.: On the efficiency of regression estimators in generalised linear models for longitudinal data. *Biometrika* **86**, 459–465 (1999)
42. Wang, N., Robins, J. M. : Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85**, 935–948 (1998)
43. White, H.A. : Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25 (1982)
44. Yi, G.Y., Cook, R.J. : Marginal methods for incomplete longitudinal data arising in clusters. *Journal of the American Statistical Association* **97(460)**, 1071–1080 (2002)
45. Zhao, L.P., Prentice, R.L. : Correlated binary regression using a quadratic exponential model. *Biometrika* **77**, 642–648 (1990)