Statistica Sinica **16**(2006), 1143-1167

LIKELIHOOD-BASED INFERENCE WITH NONIGNORABLE MISSING RESPONSES AND COVARIATES IN MODELS FOR DISCRETE LONGITUDINAL DATA

Amy L. Stubbendick and Joseph G. Ibrahim

Biogen and University of North Carolina at Chapel Hill

Abstract: We propose methods for estimating parameters in two types of models for discrete longitudinal data in the presence of nonignorable missing responses and covariates. We first present the generalized linear model with random effects, also known as the generalized linear mixed model. We specify a missing data mechanism and a missing covariate distribution and incorporate them into the complete data log-likelihood. Parameters are estimated via maximum likelihood using the Gibbs sampler and a Monte Carlo EM algorithm. The second model is a marginal model for correlated binary responses and discrete covariates with finite range, both of which may be nonignorably missing. We incorporate the missing data mechanism and the missing covariate distribution into the multivariate probit model defined by Chib and Greenberg (1998). We use the EM by method of weights (Ibrahim, 1990) and sample the latent normal variables conditional on a particular response and covariate pattern. The M-steps for each model are like a complete data maximization problem, and standard methods are used. Standard errors for the parameter estimates are computed using the multiple imputation method of Goetghebeur and Ryan (2000). We discuss the advantages and disadvantages of each model and give some guidance as to when one model might be chosen over the other. We illustrate both models using data from an environmental study of dyspnea in Chinese cotton factory workers.

Key words and phrases: Generalized linear mixed model, Gibbs sampling, Monte Carlo EM algorithm, multivariate probit model, nonignorable missing data.

1. Introduction

Correlated discrete observations arise in a variety of settings. For example, in many clinical trials, a dichotomous response to treatment is measured repeatedly over time. In toxicology studies, ordinal or count responses are often measured on animals that are correlated within litter. Various complete data methods have been proposed for modeling discrete correlated data. Two common approaches are generalized estimating equations (GEEs) (Liang and Zeger (1986)) and generalized linear models with random effects, often referred to as generalized linear mixed models (GLMMs) (see, for example, Zeger and Karim

(1991) and Breslow and Clayton (1993)). Others have used latent variable models, where the joint distribution of the binary responses is specified by relating it to the joint distribution of some underlying latent continuous responses. When the joint continuous distribution is taken to be multivariate normal, the model is referred to as a multivariate probit model (Ashford and Sowden (1970) and Ochi and Prentice (1984)).

A common problem in the analysis of this type of data involves missingness with a possibly nonignorable response mechanism. Subjects often drop out of longitudinal studies or may miss visits intermittently. When nonresponse is unrelated to the values of the missing variables and the parameters of the missing data mechanism are distinct from the response (sampling) model, the nonresponse is called ignorable, and likelihood-based methods or GEEs with slight modifications (see Robins, Rotnitzky and Zhao (1995)) can be used. However, if conditional on the observed data, nonresponse depends on the missing values, the nonresponse is termed nonignorable, and methods that do not model the missingness mechanism are subject to bias. Many likelihood-based methods have been proposed for handling nonignorably missing correlated discrete responses. Approaches using selection models include Baker (1995), Fitzmaurice, Laird and Zahner (1996), Molenberghs, Kenward and Lesaffre (1997), and Ibrahim, Chen and Lipsitz (2001). Approaches based on pattern-mixture models have been taken by Ekholm and Skinner (1998), Fitzmaurice and Laird (2000) and Birmingham and Fitzmaurice (2002). Many of these models assume monotone patterns of missingness, commonly referred to as dropout. Follmann and Wu (1995) propose a class of shared parameter models for nonignorable nonresponse that can be specified as a random effects model for the primary response, combined with a model for the missingness in which the random effects are treated as covariates. They condition on the data that describes missingness and use a conditional model for inference. Rotnitzky, Robins and Scharfstein (1998) develop a class of estimators for generalized linear models (GLMs) with nonignorable missing responses that are based on inverse probability weighted estimating equations.

The previous estimation methods have all been developed for data involving nonignorably missing discrete responses, while the covariates are assumed to be completely observed. There is less literature on maximum likelihood estimation with missing covariates. Lipsitz and Ibrahim (1996) present a conditional model for missing at random (MAR) covariates in parametric regression models, and Ibrahim, Lipsitz and Chen (1999) propose a method for estimating parameters in GLMs with missing covariates and a nonignorable missing data mechanism. Roy and Lin (2002) and Stubbendick and Ibrahim (2003) propose maximum likelihood methods for nonignorable missing responses and covariates in the normal random effects model, where the response variable is continuous. However, their methodologies rely on continuity of the response and cannot be easily extended to discrete data. Hence, none of the literature has examined maximum likelihood estimation for correlated discrete responses when both the responses and the covariates may be nonignorably missing. This is a very common occurrence in longitudinal studies, as discrete outcomes are commonly measured and the probability of missing a scheduled visit may depend on the value of both the missing response and covariates at that time point. In addition, a subject's response and covariate values can be nonignorably missing at one time point and then measured at the next, resulting in arbitrary, nonmonotone patterns of missingness in both variables.

In this paper, we propose two models for correlated discrete data and discuss estimation with nonignorable missing responses and covariates. The first model is very general and can be used for various types of discrete data when the objective is to make inferences about individuals rather than population averages. We estimate parameters in the generalized linear mixed model (GLMM) with nonignorable missing responses and covariates by specifying a missing data mechanism and a missing covariate distribution and incorporating them into the complete data log-likelihood. We use the Monte Carlo EM algorithm (MCEM) and draw samples from the joint distribution of the missing data given the observed data and current parameter estimates. The method is very general and can be used with various link functions. In addition, there are no restrictions on the covariates; they can be discrete, continuous, and time-varying. However, because the random effects are treated as missing data, the method may not be computationally feasible for models with more than one or two random effects. Also, a high degree of autocorrelation typically results when doing Gibbs sampling with random effects (see Gelfand, Sahu and Carlin (1996)). This can cause problems in the Gibbs sampler and may result in lack of convergence.

Therefore, we consider an alternative marginal model for correlated binary responses that does not involve random effects and avoids many of the issues mentioned above. We incorporate the missing data mechanism and the missing covariate distribution into the multivariate probit model defined by Chib and Greenberg (1998). We use the EM by method of weights (Ibrahim (1990)) and sample the latent normal variables conditional on a particular response and covariate pattern. We assign weights to each sample based on the probability of each pattern and sum over all possible patterns. While this model is not as general as the GLMM and the method cannot accommodate continuous covariates, it may be applicable in many situations and offers an alternative to the GLMM when one is interested in making inferences about population averages. The M-steps for each model are like a complete data maximization problem and standard methods are used. Standard errors for the parameter estimates are computed using the multiple imputation method of Goetghebeur and Ryan (2000).

The rest of the paper is organized as follows. In the next section, we discuss the GLMM and show how maximum likelihood estimation can be done in the presence of nonignorable missing response and covariate data. In Section 3, we focus on the multivariate probit model and again use maximum likelihood for estimation with nonignorable missing responses and covariates. In Section 4, we demonstrate the two models using data from a study of Chinese cotton factory workers, and in Section 5, we give some discussion of the models.

2. The Generalized Linear Mixed Model

2.1. Model and notation

The generalized linear model (GLM) with random effects, also known as the generalized linear mixed model (GLMM), is the GLM generalization of the normal linear random effects model described by Laird and Ware (1982). It is commonly defined as follows. For a given individual i with $j = 1, \ldots, n_i$ repeated measurements, outcome y_{ij} is modeled as

$$f(y_{ij}|\beta, b_i, \tau) = \exp\left[\tau\{y_{ij}\theta(\eta_{ij}) - g(\theta(\eta_{ij}))\} + c(y_{ij}, \tau)\right] \ , \tag{2.1.1}$$

where y_i is $n_i \times 1$, τ is a scalar dispersion parameter, $\theta(\cdot)$ is the link function, $\eta_{ij} = x'_{ij}\beta + z'_{ij}b_i$ is the linear predictor, β is a $p \times 1$ vector of unknown regression parameters, x'_{ij} is the jth row of the $n_i \times p$ matrix of fixed covariates X_i, and z'_{ij} is the jth row of Z_i, the $n_i \times q$ matrix of fixed covariates for the $q \times 1$ vector of random effects b_i. The link is said to be the canonical link when $\theta(\eta_{ij}) = \eta_{ij}$. Without loss of generality, we assume that $\tau = \tau_0$, where τ_0 is known, as $\tau_0 = 1$ in logistic and Poisson regression. Hence, we write $c(y, \tau_0) = c(y)$ and $f(y_{ij}|\beta, b_i, \tau_0) = f(y_{ij}|\beta, b_i)$ in (2.1.1). Furthermore, we assume throughout that $b_i \sim N_q(0, D)$, where D is a $q \times q$ unknown covariance matrix. If we have complete data, and letting $y = (y_{11}, \ldots, y_{Nn_N})'$, $X = (X'_1, \ldots, X'_N)'$, $Z = \mathrm{diag}(Z_1, \ldots, Z_N)$, and $b = (b'_1, \ldots, b'_N)'$, then the likelihood based on N subjects for the GLMM is given by

$$f(y, b|\beta, D) = \prod_{i=1}^{N}\prod_{j=1}^{n_i} f(y_{ij}|\beta, b_i)\ f(b_i|D) \ .$$

Inference is based on the marginal likelihood of (β, D) with the random effects integrated out. This is given by

$$f(y|\beta, D) = \int_{R^{N_q}} f(y, b|\beta, D)\ db \ , \tag{2.1.2}$$

where R^{N_q} denotes the N_q-dimensional Euclidean space. Thus, even with complete data, the likelihood function involves very high-dimensional integration, and in general, (2.1.2) does not have a closed form. Only in certain special cases with certain link functions can the random effects be integrated out. When some components of y and/or X are nonignorably missing, estimation based on the observed data likelihood becomes even more complex. Two additional integrations over the missing response and covariate data are needed, and a missing data mechanism, as well as a model for the covariates, must be introduced. Therefore, we present an MCEM algorithm that makes estimation of the parameters in GLMMs with nonignorable missing response and covariate data feasible.

2.2. Estimation with nonignorable missing response and covariate data

We combine methodologies put forth in previous papers and propose a method for likelihood-based inference in GLMMs with nonignorable missing response and covariate data. The method can accommodate missing discrete responses, as well as missing covariates that are either discrete or continuous, and time-varying. This extension presents numerous modeling and computational challenges. The missing data mechanism depends on both the missing responses and covariates, and the missing covariate distribution must be able to accommodate both discrete and continuous longitudinal variables. As shown above, the random effects cannot be eliminated easily in a general GLMM, so they must be accounted for in estimation. We emphasize that the model is very general and can accommodate any GLMM with nonignorable missing response and covariate data.

Under selection modeling, a parametric model for the missing data mechanism conditional on potentially missing values is incorporated into the complete data log-likelihood. As in Stubbendick and Ibrahim (2003), we let $r_i = (u_i, v_i)'$, and we define the missing data mechanism as the distribution of the $(n_i + n_i p) \times 1$ random vector r_i, where $u_{ij} = 1$ if y_{ij} is missing, 0 otherwise, and $v_{ijk} = 1$ if x_{ijk} is missing, 0 otherwise, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$, $k = 1, \ldots, p$. This distribution is indexed by the parameter vector ϕ and is a multinomial distribution with $2^{n_i + n_i p}$ cell probabilities. Hence, the complete data density for subject i is given by

$$f(y_i, X_i, b_i, r_i | \beta, \alpha, D, \phi) = f(y_i | \beta, X_i, b_i) f(X_i | \alpha) f(b_i | D) f(r_i | \phi, y_i, X_i) \ . \quad (2.2.1)$$

We assume throughout that the distributions of X_i and r_i do not depend on b_i and, in this sense, we do not consider the most general nonignorable missing data mechanisms but rather a smaller class of nonignorable missing data mechanisms in which the the probability of missingness can only depend on y_i, X_i, or both. We specify the distributions for X_i and r_i via a sequence of one-dimensional conditional distributions (see Stubbendick and Ibrahim (2003)) and approximate a correlation structure that would be induced by the inclusion of random effects. We also assume that the columns of Z_i are a subset of the fixed and observed columns of X_i, so Z_i will be suppressed when writing out conditional distributions. This is a reasonable assumption, since Z_i is usually the design on time. Hence, the complete data log-likelihood based on (2.2.1) is given by

$$l(\gamma) = \sum_{i=1}^{N} \log\left[f(y_i|\beta, X_i, b_i)\right] + \log\left[f(X_i|\alpha)\right] + \log\left[f(b_i|D)\right] + \log\left[f(r_i|\phi, y_i, X_i)\right],$$

where $\gamma = (\beta, \alpha, D, \phi)$ denotes all of the parameters. Estimation of (β, D) is of interest, with (α, ϕ) being viewed as nuisance parameters.

For ease of exposition, write $y_i = (y_{mis,i}, y_{obs,i})$, where $y_{mis,i}$ is the $s_i \times 1$ vector of missing components of y_i. Also, write $\text{Vec}(X_i) = (x_{mis,i}, x_{obs,i})$, where $x_{mis,i}$ is the $w_i \times 1$ vector of missing components of X_i, and $x_{obs,i}$ is the $(n_i p - w_i) \times 1$ vector of observed components of X_i. Assuming arbitrary and nonmonotone patterns of missing data in y_i and X_i means that some permutation of the indices of y_i and $\text{Vec}(X_i)$ can be written as $(y_{mis,i}, y_{obs,i})$ and $(x_{mis,i}, x_{obs,i})$, respectively. The E-step of the EM algorithm consists of calculating the expected value of the complete data log-likelihood given the observed data and current parameter estimates. Since b_i is also unobserved, we can view it as missing data and integrate over it in the E-step. Thus, the E-step for the ith observation at the $(t+1)^{st}$ iteration is

$$\begin{aligned}
&Q_i(\gamma|\gamma^{(t)})\\
&= E[l(\gamma; y_i, X_i, b_i, r_i)|\gamma^{(t)}, y_{obs,i}, x_{obs_i}, r_i]\\
&= \iiint \log[f(y_i|\beta, X_i, b_i)] f(b_i, y_{mis,i}, x_{mis,i}|\gamma^{(t)}, y_{obs,i}, x_{obs,i}, r_i) db_i dy_{mis,i} dx_{mis,i}\\
&\quad + \iiint \log[f(X_i|\alpha)] f(b_i, y_{mis,i}, x_{mis,i}|\gamma^{(t)}, y_{obs,i}, x_{obs,i}, r_i) db_i dy_{mis,i} dx_{mis,i}\\
&\quad + \iiint \log[f(b_i|D)] f(b_i, y_{mis,i}, x_{mis,i}|\gamma^{(t)}, y_{obs,i}, x_{obs,i}, r_i) db_i dy_{mis,i} dx_{mis,i}\\
&\quad + \iiint \log[f(r_i|\phi, y_i, X_i)] f(b_i, y_{mis,i}, x_{mis,i}|\gamma^{(t)}, y_{obs,i}, x_{obs,i}, r_i) db_i dy_{mis,i} dx_{mis,i},
\end{aligned} \tag{2.2.2}$$

where $\gamma^{(t)} = (\beta^{(t)}, \alpha^{(t)}, D^{(t)}, \phi^{(t)})$ and $f(b_i, y_{mis,i}, x_{mis,i}|\gamma^{(t)}, y_{obs,i}, x_{obs,i}, r_i)$ represents the conditional distribution of the missing data given the observed data and the current parameter estimates. Note that $f(X_i|\alpha)$ can be factored as

$f(x_{mis,i}|\alpha, x_{obs,i})f(x_{obs,i}|\alpha)$, and since $f(x_{obs,i}|\alpha)$ does not depend on any of the unobserved quantities, it is fixed and will not affect the E- or M-step. Hence, the completely observed covariates do not need to be modeled, and $f(X_i|\alpha)$ can be replaced with $f(x_{mis,i}|\alpha, x_{obs,i})$.

We can use the MCEM algorithm given by Wei and Tanner (1990) to evaluate (2.2.2) at the $(t+1)^{st}$ iteration of EM. To do this, we need to generate a sample from

$$[b_i, y_{mis,i}, x_{mis,i}|\gamma^{(t)}, y_{obs,i}, x_{obs,i}, r_i]$$

for each i. This can be accomplished using the Gibbs sampler by sampling from the following complete conditionals:

$$[b_i|\gamma^{(t)}, y_{obs,i}, y_{mis,i}, x_{obs,i}, x_{mis,i}, r_i] \propto [y_i|\beta^{(t)}, X_i, b_i]\ [b_i|D^{(t)}]\ , \qquad (2.2.3)$$

$$[y_{mis,i}|\gamma^{(t)}, y_{obs,i}, x_{obs,i}, x_{mis,i}, b_i, r_i] \propto [r_i|\phi^{(t)}, y_i, X_i]\ [y_i|\beta^{(t)}, X_i, b_i]\ , \qquad (2.2.4)$$

$$\begin{aligned}[x_{mis,i}|\gamma^{(t)}, y_{obs,i}, y_{mis,i}, x_{obs,i}, b_i, r_i] \propto {} & [r_i|\phi^{(t)}, y_i, X_i]\ [y_i|\beta^{(t)}, X_i, b_i] \\ & \times[x_{mis,i}|\alpha^{(t)}, x_{obs,i}]\ . \qquad (2.2.5)\end{aligned}$$

Note that $[r_i|\phi^{(t)}, y_i, X_i]$ will be log-concave in $x_{mis,i}$ if each $[r_{ij}|\phi^{(t)}, y_i, X_i]$ is taken to be either a logistic or probit regression model, due to the interchangeability of the covariates and regression coefficients arising from the structure of a GLM (i.e., the density for any GLM depends on the covariates only through $\eta_i = X_i\beta$, see Ibrahim, Lipsitz and Chen (1999)). For the same reason, $[y_i|\beta^{(t)}, X_i, b_i]$ will be log-concave in the components of $x_{mis,i}$ and b_i, and since $f(b_i|D^{(t)})$ is a normal density, it will be log-concave in b_i. Finally, $[x_{mis,i}|\alpha^{(t)}, x_{obs,i}]$ will be log-concave in the components of $x_{mis,i}$ if each $[x_{mis,ijk}|\alpha^{(t)}, x_{obs,i}]$ is one of many exponential family distributions. Log-concavity in the outcome variable is a property of most continuous distributions in the exponential family (see Stubbendick and Ibrahim (2003)). Thus, the products on the right side of (2.2.5) and (2.2.3) are composed of log-concave densities, and since the sum of the logs of log-concave densities is a concave function, the Gibbs sampler along with the adaptive rejection algorithm of Gilks and Wild (1992) can be used to sample from the complete conditionals. For the missing discrete responses and covariates, the log-concavity property does not apply and rejection sampling does not need to be done. One can simply sample directly from the appropriate discrete distribution.

Let $a_{i1}, \ldots, a_{im_i}$ be a sample of size m_i from the joint distribution of $[b_i, y_{mis,i}, x_{mis,i}|\gamma^{(t)}, y_{obs,i}, x_{obs,i}, r_i]$ obtained via the Gibbs sampler in conjunction with the adaptive rejection algorithm as described above. Note that each a_{ik} will be a $(q + s_i + w_i) \times 1$ vector for $k = 1, \ldots, m_i$ and that each a_{ik} depends on the iteration number which is suppressed. Also, let $b_i^{(k)}$ be a vector composed of the first q components of a_{ik}, let $y_i^{(k)} = (y_{mis,i}^{(k)}, y_{obs,i})$, where $y_{mis,i}^{(k)}$ is composed of the next s_i components of a_{ik}, and let $\mathrm{Vec}(X_i^{(k)}) = (x_{mis,i}^{(k)}, x_{obs,i})$, where $x_{mis,i}^{(k)}$ is composed of the last w_i components of a_{ik}. The E-step for the ith observation at the $(t+1)^{st}$ iteration can now be written as

$$\begin{aligned} Q_i(\gamma|\gamma^{(t)}) = \frac{1}{m_i}\sum_{k=1}^{m_i} \log\,[f(y_i^{(k)}|\beta, X_i^{(k)}, b_i^{(k)})] + \frac{1}{m_i}\sum_{k=1}^{m_i} \log\,[f(x_{mis,i}^{(k)}|\alpha, x_{obs,i})] \\ + \frac{1}{m_i}\sum_{k=1}^{m_i} \log\,[f(b_i^{(k)}|D)] + \frac{1}{m_i}\sum_{k=1}^{m_i} \log\,[f(r_i|\phi, y_i^{(k)}, X_i^{(k)})]\ , \end{aligned}$$

and the E-step for all N observations is given by $Q(\gamma|\gamma^{(t)}) = \sum_{i=1}^{N} Q_i(\gamma|\gamma^{(t)})$.

Note that for each subject, each $(b_i, y_{mis,i}, x_{mis,i})$ gets filled in by a set of m_i values, each contributing a weight of $1/m_i$. Thus, we are essentially using the EM by method of weights which was introduced by Ibrahim (1990). A slight difference here is that instead of using exact weights for discrete covariate values, we sample all missing values and approximate the weights by $\sum_1^{m_i} I(\text{covariate pattern j})/m_i$, where I is an indicator variable. We have written the E-step in its most general form. In most applications, one would set $m_i = m$ for all i. However, one can also let m vary with each EM iteration. Wei and Tanner (1990) recommend increasing m as the current approximation moves closer to the true maximizer.

The resulting M-step is now a complete data maximization problem and is straightforward to compute. The score vector is composed of the first derivatives of $Q(\gamma|\gamma^{(t)})$ with respect to each parameter in γ, and the Hessian matrix is simply the matrix of the second derivatives of $Q(\gamma|\gamma^{(t)})$. That is,

$$\begin{aligned} \dot{Q}(\gamma \mid \gamma^{(t)}) &= \sum_{i=1}^{N} \dot{Q}_i(\gamma \mid \gamma^{(t)}) = \sum_{i=1}^{N} \frac{1}{m_i} \sum_{k=1}^{m_i} \frac{\partial l(\gamma; y_{obs,i}, x_{obs,i}, a_{ik}, r_i)}{\partial \gamma}\ , \\ \ddot{Q}(\gamma \mid \gamma^{(t)}) &= \sum_{i=1}^{N} \ddot{Q}_i(\gamma \mid \gamma^{(t)}) = \sum_{i=1}^{N} \frac{1}{m_i} \sum_{k=1}^{m_i} \frac{\partial^2 l(\gamma; y_{obs,i}, x_{obs,i}, a_{ik}, r_i)}{\partial \gamma \partial \gamma'}\ . \end{aligned}$$

The complexity of estimation usually depends on the structure of D, but the estimation of D, as well as of β, α and ϕ, corresponds to a complete data maximization problem. In principle, one can use any existing complete data software to obtain the estimates.

To obtain the asymptotic covariance matrix of $\widehat{\gamma}$, the estimate of γ at EM convergence, one can theoretically use the method of Louis (1982). The matrix of second derivatives of $Q(\widehat{\gamma}|\widehat{\gamma})$ will be block diagonal in β, D, α, and ϕ, since the parameters are distinct. However, the method requires a large number of derivatives, and the observed information matrix must be inverted. Computationally, this can be extremely difficult and possibly numerically unstable, especially with a large number of nuisance parameters. In addition, some error is introduced by Gibbs sampling. For these reasons, Stubbendick and Ibrahim (2003) used a bootstrap algorithm for obtaining standard errors of the parameters. Their method, however, does not require sampling of the random effects which can be very time-consuming and requires adequate storage space. A simpler variance estimation method that can be used in the EM context has been proposed by Goetghebeur and Ryan (2000). They impute possible values for the missing data upon convergence of the EM algorithm. Each imputed data set yields naive point and variance estimates for the parameters of interest. The variance of the EM estimator can be found as a weighted sum of the mean of the imputation variances and the empirical variance of the imputation point estimates with weights 1 and $(1 + 1/b)$, respectively, where b is the number of imputations. Hence, for our model, variance estimation would proceed as follows.

(1) Obtain $\widehat{\gamma}$ by running the EM algorithm until convergence.

(2) Using $\widehat{\gamma}$, impute one value for each missing response, covariate, and random effect (i.e., take one sample from $[b_i, y_{mis,i}, x_{mis,i}|\widehat{\gamma}, y_{obs,i}, x_{obs,i}, r_i]$ after a burn-in).

(3) Obtain parameter estimates and variances based on the information matrix as if one has complete data.

(4) Repeat steps (2) and (3) b times.

(5) Obtain the final variance estimates as: (mean of the imputation variances) + (1+1/b)(empirical variance of the imputation point estimates).

3. The Multivariate Probit Model

For the general GLMM, we have shown that the MCEM algorithm is a very powerful and necessary tool for evaluating the E-step given in (2.2.2). The method, however, may not be computationally feasible if q, s_i, and/or w_i are large. In some cases, there may be a very large number of variables to sample in the Gibbs sampler. A computational explosion arises when the random effects need to be sampled in addition to the missing responses and covariates, and there may not be enough storage space on most computers. Also, sampling of the random effects may result in a high degree of autocorrelation that can cause problems in the Gibbs sampler and lack of convergence. Thus, while the model is very general, it does have some potentially major computational drawbacks.

For these reasons, we propose a marginal model for correlated binary responses that does not involve random effects and avoids many of the issues mentioned above. By marginal, we mean that the expected response is modeled conditional only on the covariates, not on other responses or the random effects. The basic premise of marginal models is to make inferences about population averages. The method, however, is only applicable to situations in which the binary outcomes can be described in terms of a correlated Gaussian distribution for latent variables that are manifested as discrete variables through a threshold specification. We also assume that the covariates are random variables that come from a discrete distribution with finite range. Nevertheless, both the responses and covariates can be nonignorably missing and may follow nonmonotone patterns of missingness. Such situations commonly arise in longitudinal studies, and this method may be preferable to the general GLMM described in Section 2.

3.1. Model and notation

Chib and Greenberg (1998) propose a convenient formulation of the multivariate probit model in terms of Gaussian latent variables. Let y_{ij} denote a binary response for the ith subject at the jth occasion, $i = 1, \ldots, N$, $j = 1, \ldots, n$, and let $y_i = (y_{i1}, \ldots, y_{in})'$ denote the collection of responses for the ith subject. Also, let $z_i = (z_{i1}, \ldots, z_{in})'$ denote an n-variate normal vector with distribution $z_i \sim N_n(X_i\beta, \Omega)$, where $X_i = \text{diag}(x'_{i1}, \ldots, x'_{in})$ is an $n \times p$ matrix of fixed discrete covariates with finite range, $p = \sum_{j=1}^{n} p_j$, $\beta = (\beta_1, \ldots, \beta_n)'$ is a $p \times 1$ unknown parameter vector, and $\Omega = \{\omega_{jk}\}$ is an unknown correlation matrix. It is important to note that Ω must be in correlation form for identifiability reasons. A parameterization in terms of covariances is not likelihood identified (see Chib and Greenberg (1998, p.348)). If we let $y_{ij} = I(z_{ij} > 0)$, $j = 1, \ldots, n$, where $I(A)$ is the indicator function of the event A, then the probability that $Y_i = y_i$ conditional on X_i, β, and Ω is given by

$$\text{pr}(Y_i = y_i|\beta, \Omega, X_i) \equiv f(y_i|\beta, \Omega, X_i) = \int_{B_{in}} \cdots \int_{B_{i1}} \phi_n(z_i|X_i\beta, \Omega)\ dz_i\ ,$$

where $\phi_n(z_i|X_i\beta, \Omega)$ is the density of an n-variate normal distribution with mean vector $X_i\beta$ and correlation matrix Ω and

$$B_{ij} = \begin{cases} (0, \infty), & \text{if } y_{ij} = 1, \\ (-\infty, 0], & \text{if } y_{ij} = 0. \end{cases}$$

Note that $B_i = (B_{i1}, \ldots, B_{in})'$ depends only on y_i and not the parameters. Chib and Greenberg (1998) use this form of the multivariate probit model and show how the MCEM algorithm can be used for maximum likelihood estimation with complete data by sampling the latent normal data.

3.2. Estimation with nonignorable missing response and covariate data

In the presence of nonignorable missing response and covariate data, we introduce a parametric model for the missing data mechanism and incorporate it into the complete data log-likelihood. Again, we let $r_i = (u_i, v_i)'$, where $u_{ij} = 1$ if y_{ij} is missing, 0 otherwise, and $v_{ijk} = 1$ if x_{ijk} is missing, 0 otherwise, $i = 1, \ldots, N$, $j = 1, \ldots, n$, $k = 1, \ldots, p$. The complete data density for subject i is given by

$$f(y_i, z_i, X_i, r_i|\beta, \Omega, \alpha, \phi) = f(y_i, z_i|\beta, \Omega, X_i)\ f(X_i|\alpha)\ f(r_i|\phi, y_i, X_i)\ .$$

Estimation of β, and possibly the parameters of Ω, is of interest with α and ϕ being viewed as nuisance parameters. We write $y_i = (y_{mis,i}, y_{obs,i})$, where $y_{mis,i}$ is the $s_i \times 1$ vector of missing components of y_i, and we write $\text{Vec}(X_i) = (x_{mis,i}, x_{obs,i})$, where $x_{mis,i}$ is the $w_i \times 1$ vector of missing components of X_i, and $x_{obs,i}$ is the $(np - w_i) \times 1$ vector of observed components of X_i. Since we have unobserved data, $(z_i, y_{mis,i}, x_{mis,i})$, we use the EM algorithm to compute the maximum likelihood estimates, and the E-step for the ith observation at the $(t+1)^{st}$ iteration is given by

$$\begin{aligned}
&Q_i(\gamma|\gamma^{(t)}) = E[l(\gamma; y_i, z_i, X_i, r_i)|\gamma^{(t)}, y_{obs,i}, x_{obs_i}, r_i]\\
=&\iiint \log[f(y_i, z_i|\beta, \Omega, X_i)] f(z_i, y_{mis,i}, x_{mis,i}|\gamma^{(t)}, y_{obs,i}, x_{obs,i}, r_i) dz_i dy_{mis,i} dx_{mis,i}\\
&+ \iiint \log[f(X_i|\alpha)] f(z_i, y_{mis,i}, x_{mis,i}|\gamma^{(t)}, y_{obs,i}, x_{obs,i}, r_i) dz_i dy_{mis,i} dx_{mis,i}\\
&+ \iiint \log[f(r_i|\phi, y_i, X_i)] f(z_i, y_{mis,i}, x_{mis,i}|\gamma^{(t)}, y_{obs,i}, x_{obs,i}, r_i) dz_i dy_{mis,i} dx_{mis,i}\\
\equiv& I_1 + I_2 + I_3\ ,
\end{aligned}$$

where $\gamma^{(t)} = (\beta^{(t)}, \Omega^{(t)}, \alpha^{(t)}, \phi^{(t)})$ and $f(z_i, y_{mis,i}, x_{mis,i}|\gamma^{(t)}, y_{obs,i}, x_{obs,i}, r_i)$ represents the conditional distribution of the missing data given the observed data and the current parameter estimates. Note that again, $f(X_i|\alpha)$ can be replaced with $f(x_{mis,i}|\alpha, x_{obs,i})$ in I_2. As in Stubbendick and Ibrahim (2003) and the previous section, we specify the distributions of $f(x_{mis,i}|\alpha, x_{obs,i})$ and $f(r_i|\phi, y_i, X_i)$ via a sequence of one-dimensional conditional distributions.

We write

$$\begin{aligned}
&f(z_i, y_{mis,i}, x_{mis,i}|\gamma^{(t)}, y_{obs,i}, x_{obs,i}, r_i)\\
&\quad = f(z_i|\gamma^{(t)}, y_i, X_i, r_i) f(y_{mis,i}, x_{mis,i}|\gamma^{(t)}, y_{obs,i}, x_{obs,i}, r_i)
\end{aligned}$$

and rewrite I_1 as

$$I_1 = \iiint \log[f(y_i, z_i|\beta, \Omega, X_i)]\ f(z_i|\gamma^{(t)}, y_i, X_i, r_i)dz_i \\ \times f(y_{mis,i}, x_{mis,i}|\gamma^{(t)}, y_{obs,i}, x_{obs,i}, r_i)dy_{mis,i}dx_{mis,i}.$$

Note that

$$\begin{aligned} f(y_i, z_i|\beta, \Omega, X_i) &= f(y_i|\beta, \Omega, X_i, z_i)\ f(z_i|\beta, \Omega, X_i) \\ &= I(z_i \in B_i)\ f(z_i|\beta, \Omega, X_i) \end{aligned}$$

so that

$$I_1 = \iiint \log[f(z_i|\beta, \Omega, X_i)]f(z_i|\gamma^{(t)}, y_i, X_i, r_i)dz_i \\ \times f(y_{mis,i}, x_{mis,i}|\gamma^{(t)}, y_{obs,i}, x_{obs,i}, r_i)dy_{mis,i}dx_{mis,i}.$$

Since $f(x_{mis,i}|\alpha, x_{obs,i})$ and $f(r_i|\phi, y_i, X_i)$ do not depend on z_i, we can easily integrate out z_i from I_2 and I_3 and rewrite these integrals as

$$I_2 = \iint \log[f(x_{mis,i}|\alpha, x_{obs,i})]\ f(y_{mis,i}, x_{mis,i}|\gamma^{(t)}, y_{obs,i}, x_{obs,i}, r_i)\ dy_{mis,i}\ dx_{mis,i}$$
$$I_3 = \iint \log[f(r_i|\phi, y_i, X_i)]\ f(y_{mis,i}, x_{mis,i}|\gamma^{(t)}, y_{obs,i}, x_{obs,i}, r_i)\ dy_{mis,i}\ dx_{mis,i}\ .$$

Let $(y^{(l)}_{mis,i}, x^{(l)}_{mis,i})$ denote response/covariate pattern l, $l = 1, \ldots, L$, and let $z_i^{(m_l)}$ denote a sample taken from

$$f(z_i|\gamma^{(t)}, y_i^{(l)}, X_i^{(l)}, r_i) = f(z_i|\gamma^{(t)}, y_i^{(l)}, X_i^{(l)})\ ,$$

where $y_i^{(l)}{=}(y^{(l)}_{mis,i}, y_{obs,i})$ and $\text{Vec}(X_i^{(l)}){=}(x^{(l)}_{mis,i}, x_{obs,i})$. From Chib and Greenberg (1998),

$$f(z_i|\gamma^{(t)}, y_i^{(l)}, X_i^{(l)}) \\ \propto \phi_n(z_i|\beta^{(t)}, \Omega^{(t)}, X_i^{(l)})\ \prod_{j=1}^{n}\{I(z_{ij} > 0)I(y_{ij}^{(l)} = 1) + I(z_{ij} \leq 0)I(y_{ij}^{(l)} = 0)\}\ .$$

This is a multivariate normal density truncated to the region specified by B_i. To sample this distribution, Geweke (1991) points out that, conditional on all other elements of z_i, the distribution of z_{ij} is a truncated normal distribution, making the Gibbs sampler desirable. The parameters of each univariate untruncated normal distribution are obtained from the usual conditional distribution formulae, and a truncated version can be simulated by the inverse distribution function method (Devroye (1986)). Geweke (1991) also presents an alternative algorithm for sampling a truncated univariate normal distribution.

The E-step for the ith observation at the $(t+1)^{st}$ iteration can now be written as

$$Q_i(\gamma|\gamma^{(t)}) = \sum_{l=1}^{L} \frac{w_{il}}{m_i} \sum_{m=1}^{m_i} \log\,[f(z_i^{(m_l)}|\beta,\Omega,X_i^{(l)})] \\ + \sum_{l=1}^{L} w_{il} \log\,[f(x_{mis,i}^{(l)}|\alpha,x_{obs,i})] + \sum_{l=1}^{L} w_{il} \log\,[f(r_i|\phi,y_i^{(l)},X_i^{(l)})]\ ,$$

where

$$w_{il} = \frac{f(r_i|\gamma^{(t)},y_i^{(l)},X_i^{(l)})f(y_i^{(l)}|\gamma^{(t)},X_i^{(l)})f(x_{mis,i}^{(l)}|\gamma^{(t)},x_{obs,i})}{\sum_{l=1}^{L} f(r_i|\gamma^{(t)},y_i^{(l)},X_i^{(l)})f(y_i^{(l)}|\gamma^{(t)},X_i^{(l)})f(x_{mis,i}^{(l)}|\gamma^{(t)},x_{obs,i})}\ .$$

The E-step for all N observations is given by $Q(\gamma|\gamma^{(t)}) = \sum_{i=1}^{N} Q_i(\gamma|\gamma^{(t)})$.

Note that the latent variables are sampled assuming a specific response/ covariate pattern. Hence, continuous covariates would make the method computationally impossible; the covariates must have a finite range. Each latent variable sample is given a weight of w_{il}/m_i, and we sum over all possible response/covariate patterns, L. Again, this is essentially the EM by method of weights (Ibrahim (1990)). We have written the E-step in its most general form. In most applications, one would set $m_i = m$ for all i, or one could also let m vary with each EM iteration.

In the M-step of the algorithm, $Q(\gamma|\gamma^{(t)})$ is maximized over γ to obtain the new parameter vector $\gamma^{(t+1)}$. For the parameters of interest, (β,Ω), we use the ECM algorithm (Meng and Rubin (1993)) and perform two conditional maximizations. Specifically, we maximize $Q(\gamma|\gamma^{(t)})$ over β, replacing Ω with $\Omega^{(t)}$. Then we replace β with $\beta^{(t+1)}$ and maximize $Q(\gamma|\gamma^{(t)})$ over Ω. The procedure for the M-step is as follows.

(i) Find $\phi^{(t+1)}$ to maximize $Q_\phi = \sum_{i=1}^{N}\sum_{l=1}^{L} w_{il} \log[f(r_i|\phi,y_i^{(l)},X_i^{(l)})]$.

(ii) Find $\alpha^{(t+1)}$ to maximize $Q_\alpha = \sum_{i=1}^{N}\sum_{l=1}^{L} w_{il} \log[f(x_{mis,i}^{(l)}|\alpha,x_{obs,i})]$.

(iii) Find $\beta^{(t+1)}$ to minimize

$$Q_\beta = \sum_{i=1}^{N}\sum_{l=1}^{L} \frac{w_{il}}{m_i} \sum_{m=1}^{m_i} (z_i^{(m_l)} - X_i^{(l)}\beta)'\Omega^{(t)-1}(z_i^{(m_l)} - X_i^{(l)}\beta)\ ,$$

which yields

$$\beta^{(t+1)} = \left(\sum_{i=1}^{N}\sum_{l=1}^{L} w_{il} X_i^{(l)'}\Omega^{(t)-1}X_i^{(l)}\right)^{-1} \left(\sum_{i=1}^{N}\sum_{l=1}^{L} \frac{w_{il}}{m_i} \sum_{m=1}^{m_i} X_i^{(l)'}\Omega^{(t)-1}z_i^{(m_l)}\right).$$

(iv) Replace $\beta^{(t)}$ with $\beta^{(t+1)}$ and find $\Omega^{(t+1)}$ to minimize

$$Q_\Omega = \sum_{i=1}^{N}\sum_{l=1}^{L}\frac{w_{il}}{m_i}\sum_{m=1}^{m_i}\frac{1}{2}\log|\Omega|+\frac{1}{2}(z_i^{(m_l)}-X_i^{(l)}\beta^{(t+1)})'\Omega^{-1}(z_i^{(m_l)}-X_i^{(l)}\beta^{(t+1)}).$$

For efficiency reasons, it would be preferable to resample z_i from $f(z_i|\gamma^{(t)}, y_i^{(l)}, X_i^{(l)})$ with $\beta^{(t)}$ replaced by $\beta^{(t+1)}$ and likewise, recalculate the weights, w_{il}. However, improvements in efficiency need to be weighted against the additional computing time that would be required.

A bootstrap algorithm may be used to obtain the covariance matrix of $\widehat{\gamma}$ (see Stubbendick and Ibrahim (2003)). Calculating the weights, however, can be computationally intensive, and due to extensive computing times, the bootstrap is not recommended. Alternatively, the method of Goetghebeur and Ryan (2000) can again be used to obtain variance estimates of the parameters of interest. Let $\widehat{\gamma}$ denote the estimate of γ at EM convergence. The estimation of the variance of $\widehat{\gamma}$ proceeds as follows.

(1) Using $\widehat{\gamma}$, impute one value for the latent variable, z_i, for each response/covariate pattern $l = 1, \ldots, L$ (i.e., take one sample from $[z_i|\widehat{\gamma}, y_i^{(l)}, X_i^{(l)}]$ after a burn-in).

(2) Calculate the L weights using $\widehat{\gamma}$.

(3) Obtain parameter estimates and variances based on the information matrix as if one has complete data. The formulas for β would be

$$\widehat{\beta} = \left(\sum_{i=1}^{N}\sum_{l=1}^{L} w_{il}X_i^{(l)'}\widehat{\Omega}^{-1}X_i^{(l)}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{l=1}^{L} w_{il}X_i^{(l)'}\widehat{\Omega}^{-1}z_i^{(imp)}\right)$$

$$\widehat{\mathrm{Var}}\,(\widehat{\beta}) = \left(\sum_{i=1}^{N}\sum_{l=1}^{L} w_{il}X_i^{(l)'}\widehat{\Omega}^{-1}X_i^{(l)}\right)^{-1},$$

where $z_i^{(imp)}$ is the imputed value for the latent variable, z_i.

(4) Repeat step (1) b times. Note that the weights will not change for different samples of the z_i's.

(5) Obtain the final variance estimates as: (mean of the imputation variances) + (1+1/b)(empirical variance of the imputation point estimates).

4. Chinese Cotton Workers Data

We demonstrate the GLMM and the multivariate probit model in the presence of nonignorable missing responses and covariates using data from an environmental study of Chinese cotton factory workers. The outcome of interest is whether or not a worker developed dyspnea, a condition involving shortness of breath and difficulty breathing. Thus, we have a discrete response defined as 1 if the worker had dyspnea, 0 otherwise. There were 912 workers examined in 1981 that were followed up in 1986 and 1992. 14.0% of the observations were missing in 1986 and 14.4% in 1992. The covariates of interest include an indicator variable for exposure to cotton dust in the factory (exposure_i, denoted 1 for exposed, 0 for not exposed), a dichotomous variable for sex (sex_i, denoted 1 for male, 0 for female), a continuous variable for height, standardized by subtracting the mean and dividing by the standard deviation (hgt_i), a continuous variable for age, standardized (age_i), a continuous variable for the number of years worked, standardized (yrswrk_{ij}), and an indicator variable for smoking status (smoke_{ij}, denoted 1 for smoker, 0 for nonsmoker). Exposure, sex, height, and age were all measured only at baseline, while number of years worked and smoking status are time-varying. If a subject missed a response in 1986 and/or 1992, there is also missing smoking status for that year. Number of years worked was imputed based on baseline values. The percentage of subjects with at least one missing observation was 23.1%, while the overall percentage of missing observations was 9.5%.

For this data, it is highly likely that both the dyspnea response and the smoking covariate are nonignorably missing. Reasons for missing a measurement may be related to both a subject's dyspnea condition and his or her smoking status. It is well known that patients sick with disease are less likely to make scheduled follow-up visits. Also, smokers are less likely to return to have measurements taken on their lungs and breathing abilities. A smoker generally does not like to be reminded about the health of his or her lungs. Since the outcome and covariate are simultaneously missing, we consider a missing data mechanism that includes indicator variables for missingness in 1986 and 1992. That is, $r_i = (u_{i2}, u_{i3})'$, where

$$u_{ij} = \begin{cases} 1, & \text{if the outcome and covariate are missing for subject } i \text{ at time } j, \\ 0, & \text{otherwise,} \end{cases}$$

and $i = 1, \ldots, 912$, $j = 2, 3$. We construct the joint distribution for the missing data indicators through a sequence of one-dimensional conditional distributions. Since the missing data may depend on both the missing response and the missing covariate, we model the missing data mechanism as

$$\begin{aligned} & f(r_i|\phi, y_i, X_i) \\ & \quad = f(u_{i2}, u_{i3}|\phi, y_i, X_i) \\ & \quad = f(u_{i3}|\phi_3, u_{i2}, y_{i3}, y_{i2}, \text{smoke}_{i3}, \text{smoke}_{i2}, \text{exposure}_i, \text{sex}_i, \text{age}_i, \text{yrswrk}_{i3}) \\ & \qquad \times f(u_{i2}|\phi_2, y_{i2}, y_{i1}, \text{smoke}_{i2}, \text{smoke}_{i1}, \text{exposure}_i, \text{sex}_i, \text{age}_i, \text{yrswrk}_{i2}). \quad (4.0.1) \end{aligned}$$

We condition the probability of observing a response/covariate in 1992 on whether or not the response/covariate was observed in 1986. We also allow the probability of response to depend on the values of the possibly missing outcome variable and covariate at the current time point, their values at the previous time point, as well as other covariates that may affect the probability of response. Each ϕ_j, $j = 2, 3$, is estimated using a logistic regression model. Other missing data mechanisms are considered in the sensitivity analyses presented in Tables 3 and 4.

We construct the joint distribution for the missing covariates also through a sequence of one-dimensional conditional distributions. We condition smoking status in 1992 on smoking status in 1986 and 1981, as well as other covariates that may affect the probability of smoking. That is,

$$\begin{aligned} & f(x_{mis,i}|\alpha, x_{obs,i}) \\ & \quad = f(\text{smoke}_{i2}, \text{smoke}_{i3}|\alpha, x_{obs,i}) \\ & \quad = f(\text{smoke}_{i3}|\alpha_3, \text{smoke}_{i2}, \text{smoke}_{i1}, \text{exposure}_i, \text{sex}_i, \text{age}_i, \text{yrswrk}_{i3}) \\ & \qquad \times f(\text{smoke}_{i2}|\alpha_2, \text{smoke}_{i1}, \text{exposure}_i, \text{sex}_i, \text{age}_i, \text{yrswrk}_{i2}) \ . \end{aligned} \tag{4.0.2}$$

The parameter vectors α_2 and α_3 are also estimated using logistic regression models. Since all other covariates are observed, they do not need to be modeled. Other missing covariate distributions are considered in the sensitivity analyses presented in Tables 5 and 6.

For the GLMM, we use the logit link and consider the random effects model at time j given by

$$\begin{aligned} \text{logit}[E(y_{ij}|b_i)] = & \beta_0 + \beta_1 \text{exposure}_i + \beta_2 \text{sex}_i + \beta_3 \text{hgt}_i + \beta_4 \text{age}_i + \beta_5 \text{yrswrk}_{ij} \\ & + \beta_6 \text{smoke}_{ij} + \beta_7 \text{time}_j + b_i + e_{ij} \ , \end{aligned} \tag{4.0.3}$$

where $\text{time}_j = (0, 5, 11)'$ for all i, $n_i = 3$ is the number of intended responses for each i, $p = 8$, $q = 1$, $\beta = (\beta_0, \ldots, \beta_7)'$, $Z_i = (1, 1, 1)'$ for all i, and b_i is a subject specific random effect with $b_i \sim N(0, \sigma_b^2)$. For the multivariate probit model, we let $z_i = (z_{i1}, \ldots, z_{i3})$ denote a tri-variate normal vector with distribution $z_i \sim N(\mu_{zi}, \Omega)$, where

$$\begin{aligned} \mu_{zij} = & \beta_0 + \beta_1 \text{exposure}_i + \beta_2 \text{sex}_i + \beta_3 \text{hgt}_i + \beta_4 \text{age}_i + \beta_5 \text{yrswrk}_{ij} + \beta_6 \text{smoke}_{ij} \\ & + \beta_7 \text{time}_j \end{aligned} \tag{4.0.4}$$

and Ω is a 3×3 unstructured correlation matrix with parameters $(\omega_{12}, \omega_{13}, \omega_{23})$. Note that alternatively, we could have used the probit link for the GLMM. However, because the expected value of y_{ij} is conditional on the random effect, the parameters in the two models would not be comparable. The multivariate probit model is a marginal model, where the parameters represent a change in the "population-averaged" response, rather than the change in any one subject's expected response.

Table 1. GLMM Maximum Likelihood Estimates (MLE) and Standard Errors (SE).

	Complete Cases			*SAS NLMIXED Procedure*			*Nonignorable Responses and Covariates*		
Variable	MLE	SE	p-value	MLE	SE	p-value	MLE	SE	p-value
Intercept	-2.430	0.178	<0.001	-2.428	0.184	<0.001	-2.456	0.151	<0.001
Exposure	0.607	0.158	<0.001	0.602	0.152	<0.001	0.503	0.132	<0.001
Sex	-0.405	0.273	0.138	-0.391	0.256	0.127	-0.461	0.234	0.049
Height	-0.097	0.115	0.399	-0.083	0.105	0.431	-0.019	0.092	0.840
Age	0.545	0.161	<0.001	0.591	0.159	<0.001	0.498	0.142	<0.001
Yrs. Worked	-0.100	0.156	0.523	-0.129	0.154	0.402	-0.055	0.139	0.693
Smoke	0.289	0.232	0.214	0.249	0.215	0.248	0.330	0.197	0.094
Time	0.005	0.015	0.728	0.007	0.014	0.601	0.055	0.013	<0.001
σ_b^2	1.028	0.072	<0.001	1.002	0.279	<0.001	0.804	0.054	<0.001

Table 1 presents the GLMM maximum likelihood estimates of $\theta = (\beta, \sigma_b^2)$ based on the models presented in (4.0.1), (4.0.2), and (4.0.3). Standard errors and p-values are also given. The estimates were obtained from the MCEM algorithm outlined in Section 2.2, and the standard errors were obtained using the method of Goetghebeur and Ryan (2000), also presented in Section 2.2. One hundred Gibbs samples were taken within each EM iteration, and the average of the samples from the previous EM iteration was used as the starting value for the current Gibbs sampler. One hundred imputations, each with a burn-in of one hundred, were used for the standard error estimation. Table 1 also includes a complete case analysis, which assumes the data are missing completely at random (MCAR), as well as estimates from the NLMIXED procedure in SAS, which is valid under MAR.

Exposure and age are highly significant in the complete case analysis, and the results from NLMIXED are similar. Results from the nonignorable model are somewhat different, however. The most striking difference can be seen in the time covariate, which is highly significant in the nonignorable model, but is insignificant in the other two models. The sex covariate is now marginally significant, and height is much less significant. The smoking covariate is slightly more significant, but inference does not change at an α-level of 0.05. The effect on the estimate and inference concerning the time covariate, however, indicates the importance of modeling the missing data mechanism and the missing covariate distribution in the presence of missing data that may be nonignorably missing.

Table 2. Multivariate Probit Maximum Likelihood Estimates (MLE) and Standard Errors (SE).

	Complete Cases			*MAR Responses and Covariates*			*Nonignorable Responses and Covariates*		
Variable	MLE	SE	p-value	MLE	SE	p-value	MLE	SE	p-value
Intercept	-1.184	0.072	<0.001	-1.197	0.063	<0.001	-1.213	0.059	<0.001
Exposure	0.272	0.069	<0.001	0.281	0.058	<0.001	0.265	0.059	<0.001
Sex	-0.197	0.113	0.080	-0.192	0.100	0.055	-0.197	0.097	0.043
Height	-0.051	0.047	0.275	-0.038	0.040	0.348	-0.028	0.040	0.490
Age	0.274	0.073	<0.001	0.295	0.065	<0.001	0.276	0.064	<0.001
Yrs. Worked	-0.060	0.068	0.377	-0.069	0.065	0.286	-0.060	0.063	0.342
Smoke	0.146	0.088	0.096	0.111	0.080	0.165	0.115	0.079	0.143
Time	0.001	0.006	0.907	0.002	0.004	0.610	0.010	0.004	0.013

Table 2 reports maximum likelihood estimates of β from the multivariate probit model based on the models presented in (4.0.1), (4.0.2) and (4.0.4). The estimates were obtained using the EM by method of weights procedure presented in Section 3.2, and the standard errors were obtained using the method of Goetghebeur and Ryan (2000), also presented in Section 3.2. One hundred Gibbs samples of the latent variables were taken for each response/covariate pattern within each EM iteration. Again, the average of the samples from the previous EM iteration was used as the starting value for the current Gibbs sampler. The latent variables were not resampled for estimation of Ω, the 3×3 correlation matrix. One hundred imputations, each with a burn-in of one hundred, were used for the standard error estimation. Table 2 also includes a complete case analysis, as well as an analysis that assumes the data are MAR.

In the complete case analysis, exposure and age are highly significant, while gender is marginally significant. Results are similar under the assumption of MAR missingness. The nonignorable model shows some striking differences. Exposure, sex, age, and time are all significant. Again, differing estimates and inferences in the nonignorable model gives some indication that the missing data in this study may be nonignorably missing. We note here that caution must be used when interpreting results from nonigonrable missing data models. The parametric form of the assumed missing data mechanism itself is not "testable" from the data, and thus the nonignorable modeling considered here can be viewed as a sensitivity analysis concerning a more complicated model. Therefore, although a model may have "passed" the tests for a certain missing data mechanism, this does not mean that one has captured, even approximately, the correct missing data mechanism. Further evidence for or against a specific nonignorable missing data mechanism typically must come from external sources of information about the data. Thus, it is very important to address the sensitivity of the modeling scheme to both the specification of the missing data mechanism and the missing covariate distribution. With this in mind, we conducted sensitivity analyses for both aspects of the model.

For the missing data mechanism sensitivity analysis, we used the covariate distribution presented in (4.0.2) and varied the missing data mechanism. We considered several different parameterizations for the missing data mechanism:

- MDM1: $f(u_{i3}|\phi_3, u_{i2}, y_{i3}, y_{i2}, \text{exposure}_i, \text{sex}_i, \text{age}_i, \text{yrswrk}_{i3})$
  $\times f(u_{i2}|\phi_2, y_{i2}, y_{i1}, \text{exposure}_i, \text{sex}_i, \text{age}_i, \text{yrswrk}_{i2})$
- MDM2: $f(u_{i3}|\phi_3, y_{i3}, y_{i2}, \text{smoke}_{i3}, \text{smoke}_{i2}, \text{exposure}_i, \text{sex}_i, \text{age}_i, \text{yrswrk}_{i3})$
  $\times f(u_{i2}|\phi_2, y_{i2}, y_{i1}, \text{smoke}_{i2}, \text{smoke}_{i1}, \text{exposure}_i, \text{sex}_i, \text{age}_i, \text{yrswrk}_{i2})$
- MDM3: $f(u_{i3}|\phi_3, u_{i2}, y_{i3}, y_{i2}, \text{smoke}_{i3}, \text{smoke}_{i2})$
  $\times f(u_{i2}|\phi_2, y_{i2}, y_{i1}, \text{smoke}_{i2}, \text{smoke}_{i1})$
- MDM4: $f(u_{i3}|\phi_3, u_{i2}, y_{i3}, \text{smoke}_{i3})$
  $\times f(u_{i2}|\phi_2, y_{i2}, \text{smoke}_{i2})$
- MDM5: $f(u_{i3}|\phi_3, u_{i2}, y_{i3}, \text{smoke}_{i3}, \text{exposure}_i, y_{i3} \times \text{exposure}_i,$
  $\text{smoke}_{i3} \times \text{exposure}_i) \times f(u_{i2}|\phi_2, y_{i2}, \text{smoke}_{i2}, \text{exposure}_i,$
  $y_{i2} \times \text{exposure}_i, \text{smoke}_{i2} \times \text{exposure}_i)$.

Results from the sensitivity analysis are presented in Tables 3 and 4. In Table 3, we see that the GLMM model is generally robust to variations in the specification of the missing data mechanism. The estimates of the exposure effect are similar, and the models all agree that exposure to cotton dust has a significant effect on the probability that an individual will develop dyspnea. In addition, all models agree that age and time are significantly associated with the outcome, while smoking has no significant effect. Inference concerning gender at an α-level of 0.05 does vary depending on the missing data mechanism. However, all are marginally significant (at α=0.10), so there is some indication that females have a higher risk of developing dyspnea. Table 4 shows that the multivariate probit model is less robust to the specification of the missing data mechanism. While all of the models agree that exposure to cotton dust and advanced age significantly increase the risk of developing dyspnea, inference concerning gender and time changes. For the original missing data mechanism, MD1, and MD2, the probability of developing dyspnea significantly increases with time, whereas MD3, MD4, and MD5 indicate there is no difference over time. Thus, we judge the effect of time with caution. The results of our sensitivity analysis for the missing data mechanism highlight the importance of gathering information on the reasons for missingness and considering a variety of models.

We also conducted a sensitivity analysis of the missing covariate distribution. In this case, we used the original missing data mechanism presented in (4.0.1) and varied the parameterization of the missing covariate distribution. The following models were considered.

Table 3. GLMM estimates and (standard errors) from missing data mechanism sensitivity analysis.

	Original					
	MDM	MDM1	MDM2	MDM3	MDM4	MDM5
Intercept	-2.456^*	-2.461^*	-2.477^*	-2.466^*	-2.456^*	-2.516^*
	(0.151)	(0.155)	(0.150)	(0.152)	(0.154)	(0.159)
Exposure	0.502^*	0.499^*	0.458^*	0.501^*	0.528^*	0.558^*
	(0.132)	(0.136)	(0.130)	(0.136)	(0.141)	(0.139)
Sex	-0.461^*	-0.419	-0.482^*	-0.446	-0.404	-0.401
	(0.234)	(0.237)	(0.228)	(0.234)	(0.235)	(0.236)
Height	-0.019	-0.022	-0.007	-0.039	-0.050	-0.044
	(0.092)	(0.095)	(0.092)	(0.097)	(0.096)	(0.097)
Age	0.498^*	0.493^*	0.473^*	0.500^*	0.538^*	0.508^*
	(0.142)	(0.140)	(0.137)	(0.148)	(0.142)	(0.148)
Yrs. Worked	-0.055	-0.047	-0.034	-0.058	-0.095	-0.065
	(0.139)	(0.139)	(0.135)	(0.144)	(0.138)	(0.144)
Smoke	0.330	0.261	0.327	0.322	0.270	0.281
	(0.197)	(0.203)	(0.197)	(0.197)	(0.198)	(0.198)
Time	0.055^*	0.056^*	0.075^*	0.047^*	0.033^*	0.041^*
	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)	(0.014)
σ_b^2	0.804^*	0.837^*	0.856^*	0.875^*	0.853^*	0.933^*
	(0.054)	(0.052)	(0.057)	(0.054)	(0.053)	(0.062)

$^*p < 0.05$

Table 4. Multivariate Probit Estimates and (Standard Errors) from Missing Data Mechanism Sensitivity Analysis.

	Original					
	MDM	MDM1	MDM2	MDM3	MDM4	MDM5
Intercept	-1.213^*	-1.222^*	-1.199^*	-1.210^*	-1.216^*	-1.216^*
	(0.059)	(0.062)	(0.061)	(0.060)	(0.063)	(0.060)
Exposure	0.265^*	0.264^*	0.256^*	0.262^*	0.273^*	0.270^*
	(0.059)	(0.056)	(0.057)	(0.059)	(0.059)	(0.059)
Sex	-0.197^*	-0.192	-0.197	-0.196^*	-0.184	-0.194
	(0.097)	(0.101)	(0.100)	(0.098)	(0.100)	(0.100)
Height	-0.028	-0.026	-0.018	-0.029	-0.038	-0.033
	(0.040)	(0.042)	(0.040)	(0.040)	(0.041)	(0.042)
Age	0.276^*	0.279^*	0.277^*	0.276^*	0.279^*	0.278^*
	(0.064)	(0.063)	(0.064)	(0.065)	(0.067)	(0.066)
Yrs. Worked	-0.060	-0.062	-0.062	-0.065	-0.068	-0.066
	(0.063)	(0.061)	(0.062)	(0.063)	(0.068)	(0.064)
Smoke	0.115	0.107	0.106	0.112	0.110	0.116
	(0.079)	(0.079)	(0.080)	(0.078)	(0.080)	(0.079)
Time	0.010^*	0.010^*	0.011^*	0.007	0.003	0.005
	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)

$^*p < 0.05$

- CD1: $f(\text{smoke}_{i3}|\alpha_3, \text{exposure}_i, \text{sex}_i, \text{age}_i, \text{yrswrk}_{i3})$
  $\times f(\text{smoke}_{i2}|\alpha_2, \text{exposure}_i, \text{sex}_i, \text{age}_i, \text{yrswrk}_{i2})$
- CD2: $f(\text{smoke}_{i3}|\alpha_3, \text{smoke}_{i2}, \text{smoke}_{i1})$
  $\times f(\text{smoke}_{i2}|\alpha_2, \text{smoke}_{i1})$
- CD3: $f(\text{smoke}_{i3}|\alpha_3, \text{smoke}_{i2}, \text{exposure}_i, \text{smoke}_{i2} \times \text{exposure}_i)$
  $\times f(\text{smoke}_{i2}|\alpha_2, \text{smoke}_{i1}, \text{exposure}_i, \text{smoke}_{i1} \times \text{exposure}_i)$.

Results from this sensitivity analysis are presented in Tables 5 and 6. Again, all of the models agree that exposure to cotton dust and advanced age are significantly associated with a greater risk of developing dyspnea. For the GLMM, inference concerning gender is different for CD2 and CD3, and inference concerning smoking status is different for CD1. When the observed covariates are not included in the models for the missing smoke covariates, or when we assume a smoking/exposure interaction, gender is no longer significantly associated with the outcome. On the other hand, when the missing smoke covariates are assumed to be independent, smoking status is significant. For the multivariate probit model, all of the models agree that females have a higher risk of developing dyspnea, but again, inference changes for the smoking covariate under CD1. The largest differences between the models, however, can be seen in the time covariate. Not only do inferences change, but the estimated direction of effect is

Table 5. GLMM estimates and (standard errors) from covariate distribution sensitivity analysis.

	Original CD	CD1	CD2	CD3
Intercept	-2.456^* (0.151)	-2.462^* (0.154)	-2.477^* (0.152)	-2.465^* (0.150)
Exposure	0.502^* (0.132)	0.504^* (0.138)	0.518^* (0.134)	0.510^* (0.133)
Sex	-0.461^* (0.234)	-0.636^* (0.234)	-0.399 (0.233)	-0.442 (0.227)
Height	-0.019 (0.092)	-0.036 (0.094)	-0.030 (0.095)	-0.032 (0.093)
Age	0.498^* (0.142)	0.523^* (0.145)	0.498^* (0.140)	0.497^* (0.139)
Yrs. Worked	-0.055 (0.139)	-0.066 (0.139)	-0.058 (0.135)	-0.057 (0.136)
Smoke	0.330 (0.197)	0.599^* (0.201)	0.276 (0.195)	0.318 (0.193)
Time	0.055^* (0.013)	0.052^* (0.013)	0.052^* (0.013)	0.053^* (0.013)
σ_b^2	0.804^* (0.054)	0.827^* (0.055)	0.811^* (0.054)	0.799^* (0.051)

$^*p < 0.05$

Table 6. Multivariate probit estimates and (standard errors) from covariate distribution sensitivity analysis.

	Original CD	CD1	CD2	CD3
Intercept	-1.213^* (0.059)	-1.227^* (0.060)	-1.222^* (0.061)	-1.225^* (0.062)
Exposure	0.265^* (0.059)	0.270^* (0.058)	0.277^* (0.058)	0.275^* (0.058)
Sex	-0.197^* (0.097)	-0.236^* (0.095)	-0.197^* (0.098)	-0.199^* (0.099)
Height	−0.028 (0.040)	−0.047 (0.041)	−0.041 (0.041)	−0.038 (0.042)
Age	0.276^* (0.064)	0.292^* (0.066)	0.286^* (0.067)	0.282^* (0.067)
Yrs. Worked	−0.060 (0.063)	−0.079 (0.066)	−0.077 (0.066)	−0.072 (0.065)
Smoke	0.115 (0.079)	0.236^* (0.080)	0.148 (0.084)	0.147 (0.078)
Time	0.010^* (0.004)	−0.003 (0.004)	−0.000 (0.004)	0.001 (0.004)

$^*p < 0.05$

also different. Once should exercise caution when interpreting this variable in the multivariate probit model.

The convergence criterion for the EM algorithm was that the distance between the tth and the $(t + 5)$th iteration for the β parameters was less than 10^{-2}. Since the random effects or the latent normal variables must be sampled for each subject, a more stringent criterion will usually not be possible due to Gibbs sampling variation. A Gibbs sample size of 1,000 per EM iteration was used to check the sensitivity of the parameter estimates to the choice of the Gibbs sample size, and we also considered a burn-in of 100 with a Gibbs sample size of 100 per EM iteration. Estimates from these runs were very similar to those using a Gibbs sample size of 100 with no burn-in (results not shown). In addition, we monitored convergence of the Gibbs sampler for the random effect by examining the Gelman and Rubin (1992) scale reduction factors and autocorrelations for a few subjects at various EM iterations. With a Gibbs sample size of 100 (no burn-in), all lag-50 autocorrelations were less than ±0.2, and the quantiles of the scale reduction factors were all near one. Realistically, convergence of the Gibbs sampler for all random effects cannot be monitored for each subject at each EM iteration. However, lack of convergence in the Gibbs sampler will most likely cause lack of convergence in the MCEM algorithm, and any serious problems should present themselves.

5. Discussion

We have proposed methods of estimation for the GLMM and the multivariate probit model in the presence of nonignorable missing response and covariate data. The GLMM is very general and can be used for various types of discrete response data when the objective is to make inferences about individuals rather than population averages. The missing covariates can be either continuous or discrete, and time-varying. The drawbacks to this model involve the random effects. Gibbs sampling of the random effects may be computationally intensive, and a high degree of autocorrelation may cause problems with convergence. Also, it is highly likely that some models may not be identifiable for certain values of q. Depending on the missing data pattern and the assumed model for the missing data mechanism, certain dimensions of q may result in a very flat likelihood and hence, a nearly nonidentified model. In addition, parameters from a marginal model may be of interest.

As an alternative, we also present a method of estimating parameters in the multivariate probit model with nonignorable missing responses and covariates. This model is appropriate for correlated binary responses and discrete covariates with finite range. The major drawback to this model, beyond its restrictions on the type of response and covariate data, is the amount of computing time required to calculate the weights. For our example, the weights involved computation of a 3-dimensional integral, and most of the models took over 30 hours to achieve convergence. Nevertheless, this model offers an alternative to the GLMM and avoids many of the issues that arise when doing Gibbs sampling with random effects.

We also point out that the methods use a selection modeling approach, where the probability of missingness is conditioned on the potentially missing responses and/or covariates. Theoretical verification of model identifiability for these types of models is a topic of further research. Additional research on Gibbs sampling with random effects and nonidentifiability in mixed models with missing discrete data is also warranted. In addition, sensitivity analyses for these models cannot be done by simply varying certain parameters in the missing data mechanism or the missing covariate distribution. As our example shows, it is important to consider a variety of models and examine how inferences may be affected. We reiterate that it is important to collect as much information as possible about the reasons for missingness, so that each model can be given appropriate consideration.

References

Ashford, J. R. and Sowden, R. R. (1970). Multi-variate probit analysis. *Biometrics* **26**, 535-546.

Baker, S. G. (1995). Marginal regression for repeated binary data with outcome subject to non-ignorable non-response. *Biometrics* **51**, 1042-1052.

Birmingham, J. and Fitzmaurice, G. M. (2002). A pattern-mixture model for longitudinal binary responses with nonignorable nonresponse. *Biometrics* **58**, 989-996.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**, 9-25.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347-361.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer Verlag, New York.

Ekholm, A. and Skinner, C. (1998). The Muscatine children's obesity data reanalysed using pattern mixture models. *Appl. Statist.* **47**, 251-263.

Fitzmaurice, G. M. and Laird, N. M. (2000). Generalized linear mixture models for handling nonignorable dropouts in longitudinal studies. *Biostatistics* **1**, 141-156.

Fitzmaurice, G. M., Laird, N. M. and Zahner, G. E. P. (1996). Multivariate logistic models for incomplete binary responses. *J. Amer. Statist. Assoc.* **91**, 99-108.

Follmann, D. and Wu, M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics* **51**, 151-168.

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1996). Efficient parametrizations for generalized linear mixed models (with discussion). In *Bayesian Statistics* bf 5 (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 165-180. Oxford University Press.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7**, 457-511

Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints. In *Computing Science and Statistics: Proceedings of the* 23*rd Symposium on the Interface* (Edited by E. Keramidas and S. Kaufman), 571-578. Fairfax Station, VA: Interface Foundation of North America.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.* **41**, 337-348.

Goetghebeur, E. and Ryan, L. (2000). Semiparametric regression analysis of interval-censored data. *Biometrics* **56**, 1139-1144.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *J. Amer. Statist. Assoc.* **85**, 765-769.

Ibrahim, J. G., Chen, M.-H. and Lipsitz, S. R. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika* **88**, 551-564.

Ibrahim, J. G., Lipsitz, S. R. and Chen, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *J. Roy. Statist. Soc. Ser. B* **61**, 173-190.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963-974.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Lipsitz, S. R. and Ibrahim, J. G. (1996). A conditional model for incomplete covariates in parametric regression models. *Biometrika* **83**, 916-922.

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44**, 226-233.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80**, 267-278.

Molenberghs, G., Kenward, M. G. and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika* **84**, 33-44.

Ochi, Y. and Prentice, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika* **71**, 531-543.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90**, 106-121.

Rotnitzky, A., Robins, J. M. and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J. Amer. Statist. Assoc.* **93**, 1321-1339.

Roy, J. and Lin, X. (2002). Analysis of multivariate longitudinal outcomes with nonignorable dropouts and missing covariates: changes in methadone treatment practices. *J. Amer. Statist. Assoc.* **97**, 40-52.

Stubbendick, A. L. and Ibrahim, J. G. (2003). Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics*, **59**, 1140-1150.

Wei, G. C. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85**, 699-704.

Zeger, S. L. and Karim, R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *J. Amer. Statist. Assoc.* **86**, 79-86.

Biogen, 14 Cambridge Center, Cambridge, MA 02142, U.S.A.

E-mail: amy_stubbendick@biogen.com

Department of Biostatistics, University of North Carolina School of Public Health, McGavran Greenberg Hall, Chapel Hill, North Carolina 27599, U.S.A.

E-mail: ibrahim@bios.unc.edu

(Received December 2004; accepted April 2005)