Journal of Data Science **5**(2007), 1-21

Pseudo-likelihood Methods for the Analysis of Longitudinal Binary Data Subject to Nonignorable Non-monotone Missingness

Michael Parzen[1], Stuart R. Lipsitz[2], Garrett M. Fitzmaurice[3], Joseph G. Ibrahim[4], Andrea Troxel[5] and Geert Molenberghs[6]
[1]*Emory University,* [2]*Medical University of South Carolina,*
[3]*Harvard School of Public Health,* [4]*University of North Carolina*
[5]*Columbia University School of Public Health and* [6]*Limburgs Universitair*

Abstract: For longitudinal binary data with non-monotone non-ignorable missing outcomes over time, a full likelihood approach is complicated algebraically, and maximum likelihood estimation can be computationally prohibitive with many times of follow-up. We propose pseudo-likelihoods to estimate the covariate effects on the marginal probabilities of the outcomes, in addition to the association parameters and missingness parameters. The pseudo-likelihood requires specification of the distribution for the data at all pairs of times on the same subject, but makes no assumptions about the joint distribution of the data at three or more times on the same subject, so the method can be considered semi-parametric. If using maximum likelihood, the full likelihood must be correctly specified in order to obtain consistent estimates. We show in simulations that our proposed pseudo-likelihood produces a more efficient estimate of the regression parameters than the pseudo-likelihood for non-ignorable missingness proposed by Troxel *et al.* (1998). Application to data from the Six Cities study (Ware, et.al, 1984), a longitudinal study of the health effects of air pollution, is discussed.

Key words: Incomplete data, maximum likelihood, repeated measurements.

1. Introduction

Longitudinal studies in which each subject is to be observed at a fixed number of time points have become very popular in social science and medical applications. For example, longitudinal data are often collected in cardiovascular, cancer and aids clinical trials and also in observational studies of chronic conditions such as arthritis and respiratory disease. We focus on the case where the response variable over time is binary (e.g. success or failure) and we are interested in modeling the marginal means or success probabilities; also, the time dependence or association between pairs of responses is commonly modeled in terms of pairwise correlations or odds ratios. In this paper we consider statistical methods for the analysis of such data when the outcome is not observed at all times. We also focus on non-ignorable missing data mechanisms (Little and Rubin, 1987), in which the probability that an outcome is missing at a given time can depend on the possibly missing value of the outcome at that time. The missing outcome data must be properly accounted for in the analysis. In general, an individual's response can be missing at one follow-up time, and be measured at the next follow-up time, resulting in a large class of distinct missingness patterns, often called "non-monotone" missingness.

To formulate a full likelihood for non-ignorable non-monotone missing outcomes over time, one must specify a model for the repeated binary outcomes of interest, and a model for the missingness mechanism. To estimate the parameters, a full likelihood approach has many nuisance parameters, is complicated algebraically, and maximum likelihood estimation can be computationally prohibitive, especially when the number of times is large. We propose a 'pseudo-likelihood' (Gong and Samaniego, 1981; Liang and Self, 1996) to estimate the covariate effects on the marginal probabilities of the outcomes, in addition to the association parameters and missingness parameters. The pseudo-likelihood requires only partial specification of the distribution of observations and missingness indicators at pairs of times, and can be much less computationally prohibitive than maximum likelihood.

In section 5, the methods proposed are applied to the Six Cities study (Ware, et.al, 1984), a longitudinal study of the health effects of air pollution. The repeated binary response is the wheezing status (0=no wheeze, 1=wheeze) of a child at ages 9, 10, 11, and 12. The covariates of interest include the child's age, maternal smoking at baseline (just before the wheeze measurement at age 9) in cigarettes per day, and the city where the child resides (Kingston-Harriman or Portage, two of the participating cities). Table 1 shows data from 25 of the 3331 subjects on file. We see from Table 1 that there is much missing data. The covariates maternal smoking and city were observed for all children. While children

were to come for a doctor visit once a year to have the respiratory status checked (including wheeze), their compliance was not compulsory; they did sometimes miss an appointment. Even after missing an appointment, however, the children often came back the following year for a doctor's visit, leading to non-monotone missing data. The percentage missing wheeze at a given time was between 30% and 40%. At age 9, 38.2% of the children have wheeze missing; at age 10, 30.5% of the children have wheeze missing; at age 11, 32.4% of the children have wheeze missing; and at age 12, 39.7% of the children have wheeze missing. In the case of wheeze, it is quite plausible that a child might miss a visit because he or she is not wheezing, and did not come in for a doctor's visit; we would expect that the parent of a child who is wheezing would be more likely to keep the doctor's appointment, and thus have wheeze measured at that time point. This implies that missingness in this study may depend on the unobserved outcome of interest and thus may be "nonignorable."

Fitzmaurice, Molenberghs, and Lipsitz (1995) have discussed longitudinal binary data with non-ignorable dropout, summarizing likelihood approaches. Baker (1995) has also discussed likelihood approaches for repeated binary measurements with nonignorable non-response, proposing models for marginal probabilities and the missingness mechanism. Diggle and Kenward (1994) and Ibrahim *et al.* (2001) have proposed likelihood based methods for longitudinal Gaussian data with nonignorable dropout. These likelihoods are formed by summing over the possible values for the unobserved responses. In contrast, Troxel *et al.* (1998) proposed a pseudo-likelihood that is formed by naively assuming that the longitudinal binary measurements are independent over time. Specifically, their pseudo-likelihood assumes a marginal logistic regression model for the outcome at each time point, and also that the missingness probability at a given time depends only on the possibly missing response at that time and the covariates (the covariates are assumed to be fully observed). The chief attraction of this pseudo-likelihood approach is that it significantly eases the numerical complexities of the full likelihood approach by reducing high-dimensional sums to sums of a single dimension. Further, it alleviates the need to specify and estimate many nuisance parameters that are needed in a full likelihood approach. In addition, asymptotically unbiased estimators of the regression parameters and missingness parameters can be obtained. However, by naively assuming independence of repeated measures across measurement occasions, their method can be highly inefficient for estimating the regression parameters, the usual target of inference. For example, results from table 1 of the paper by Troxel et al (1998) indicate that their pseudo-likelihood method can be very inefficient compared to the MLE. Thus, in this paper, we propose an alternative pseudo-likelihood for non-ignorably missing data that yields more efficient estimates than the pseudo-likelihood proposed

by Troxel at al (1998). In particular, we propose a pseudo-likelihood approach, based on specifying the distribution of the data at all pairs of times on the same subject; our pseudo-likelihood makes no assumptions about the joint distribution of the data at three or more times on the same subject, so the method can be considered semi-parametric in this sense. Compared to maximum likelihood, which requires the full likelihood to be correctly specified in order to obtain consistent estimates, the pseudo- likelihood estimates are consistent as long as the bivariate distributions are correctly specified.

We show in simulations in Section 6 that our proposed pseudo-likelihood produces a more efficient estimate of the regression parameters than the pseudo-likelihood of Troxel *et al.* (1998). The issues of estimability and non-identifiability arises often in nonignorable non-response models as pointed out by Baker and Laird (1988). Our method also produces estimates of the association parameters, which are not available with the method of Troxel *et al.* (1998). In Section 3, we review the pseudo-likelihood of Troxel at al., and in Section 4, we outline our pseudo-likelihood. Section 5 illustrates the methods with the Six Cities example. In Section 6, we present results from our simulation study.

2. Underlying Data Model

We assume that n independent subjects are to be observed at T occasions. For the i^{th} individual ($i = 1, ..., n$), we can form a $T \times 1$ vector, $\mathbf{Y}_i = [Y_{i1}, ..., Y_{iT}]'$, where the binary random variable Y_{it} equals 1 if the i^{th} individual has response 1 (say "success") at time t, and 0 otherwise. Each individual also has a $J \times 1$ covariate vector $\mathbf{x}_i$. We assume that all covariates are time-stationary and are fully observed. With a binary response measured at each occasion, there are 2^T possible response sequences over time, and $\mathbf{Y}_i$ has a multinomial distribution with 2^T joint cell probabilities ($2^T - 1$ of which are non-redundant),

$$p_{iy_1 \ldots y_T}(\mathbf{x}_i, \boldsymbol{\theta}) = f_y(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}) = \text{pr}\{Y_{i1} = y_1, ..., Y_{iT} = y_T|\mathbf{x}_i, \boldsymbol{\theta}\}, \tag{2.1}$$

parameterized by $\boldsymbol{\theta}$. The models we consider are often referred to as marginal regression models and they describe the expected value of an individual's binary response at time t, or, equivalently, the probability of success at time t. If we denote the regression parameters relating the probability of success to $\mathbf{x}_i$ by $\boldsymbol{\beta}$, then we can partition the parameter vector $\boldsymbol{\theta}$ as $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \boldsymbol{\alpha}')$, where $\boldsymbol{\alpha}$ contains the association parameters between pairs of variables Y_{is} and Y_{it}. Finally, let p_{it} be the marginal probability of success at occasion t for the i^{th} individual, obtained by summing the cell probabilities over all but the t^{th} subscript, which is set equal to 1,

$$p_{it} = \text{pr}(Y_{it} = 1|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}) = E(Y_{it}|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}) = p_{i+\ldots+1+\ldots+}(\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}). \tag{2.2}$$

We consider models for the distribution of $\mathbf{Y}_i$ such that p_{it} in (2.2) follows a logistic regression model with parameter vector $\boldsymbol{\beta}$,

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = \mathbf{x}_{it}'\boldsymbol{\beta}\ , \tag{2.3}$$

where $\mathbf{x}_{it}$ contains $\mathbf{x}_i$ in addition to time trends. The marginal distribution of Y_{it} given $\mathbf{x}_i$ follows a Bernoulli distribution,

$$f(y_{it}|\mathbf{x}_i,\boldsymbol{\beta}) = p_{it}^{y_{it}}(1-p_{it})^{(1-y_{it})}\ . \tag{2.4}$$

There are many different multinomial models for the distribution of $\mathbf{Y}_i$ such that (2.3) holds, including multinomial models where the association parameters are correlations (Bahadur, 1961) or odds ratios (Molenberghs and Lesaffre, 1994).

In many longitudinal studies, individuals are often not observed at all T occasions on account of some stochastic missing data mechanism. Then, it is convenient to introduce a $T \times 1$ random vector for the i^{th} individual, $\mathbf{R}_i$, whose t^{th} component, R_{it}, equals 1 if Y_{it} is observed, and 0 if Y_{it} is missing. The full data for the i^{th} individual are given by $\mathbf{Y}_i$ and $\mathbf{R}_i$, with joint distribution

$$f_{y,r}(\mathbf{y}_i,\mathbf{r}_i|\mathbf{x}_i,\boldsymbol{\theta},\boldsymbol{\gamma}) = f_r(\mathbf{r}_i|\mathbf{y}_i,\mathbf{x}_i,\boldsymbol{\gamma})f_y(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta});$$

$f_r(\mathbf{r}_i|\mathbf{y}_i,\mathbf{x}_i,\boldsymbol{\gamma})$ is referred to as the "missing data mechanism", and is indexed by the parameter vector $\boldsymbol{\gamma}$.

Next, we briefly describe some possible missing data mechanisms. First, we partition $\mathbf{Y}_i$ into the observed components, $\mathbf{Y}_i^o$, and the unobserved components $\mathbf{Y}_i^u$. If the data are *missing completely at random* (MCAR), then

$$f_r(\mathbf{r}_i|\mathbf{y}_i,\mathbf{x}_i,\boldsymbol{\gamma}) = f_r(\mathbf{r}_i|\mathbf{x}_i,\boldsymbol{\gamma})$$

which does not depend on any observed or unobserved components of $\mathbf{Y}_i$. If the data are *missing at random* (MAR), then

$$f_r(\mathbf{r}_i|\mathbf{y}_i,\mathbf{x}_i,\boldsymbol{\gamma}) = f_r(\mathbf{r}_i|\mathbf{y}_i^o,\mathbf{x}_i,\boldsymbol{\gamma}),$$

can depend on any observed components of $\mathbf{Y}_i$. Both missing at random and missing completely at random fall within the class of *ignorable* missing data mechanisms (with the added provision that $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are disjoint, in the sense that their parameter spaces are separable). If the missing data are non-ignorable, then

$$f_r(\mathbf{r}_i|\mathbf{y}_i,\mathbf{x}_i,\boldsymbol{\gamma}) = f_r(\mathbf{r}_i|\mathbf{y}_i^o,\mathbf{y}_i^u,\mathbf{x}_i,\boldsymbol{\gamma}),$$

can depend on any observed or unobserved components of $\mathbf{Y}_i$. Finally, we assume missingness is not necessarily monotone, e.g., responses of patients can be missing

at one study visit, and then can be obtained at future visits. The observed data are $\mathbf{R}_i$ and $\mathbf{Y}_i^o$, with distribution

$$f_{y^o,r}(\mathbf{y}_i^o, \mathbf{r}_i|\mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \sum_{\mathbf{y}_i^u} f_r(\mathbf{r}_i|\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\gamma}) f_y(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}), \tag{2.5}$$

where the summation is over all possible values of the unobserved data, $\mathbf{Y}_i^u$. The *observed data likelihood* is determined by $f_{y^o,r}(\mathbf{y}_i^o, \mathbf{r}_i|\mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\gamma})$. The full specification of (2.5) involves complex multinomial distributions for both $f_r(\mathbf{r}_i|\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\gamma})$ and $f_y(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta})$. With T large, in order to estimate $\boldsymbol{\beta}$ one has to specify a likelihood with many nuisance parameters. To alleviate the need to specify the full likelihood, we propose a pseudo-likelihood approach. In the next section, we briefly discuss the pseudo-likelihood of Troxel *et al.*; then, in the following section, we discuss our proposed pseudo-likelihood.

3. Pseudo-Likelihood under Naive Assumption of Independence

In this section we review the pseudo-likelihood approach proposed by Troxel et al (1998) that naively assumes independence across measurement occasions. The resulting pseudo-likelihood is a product of simple marginal terms and can be used to estimate the marginal regression parameters $\boldsymbol{\beta}$ and the marginal missingness parameters $\boldsymbol{\gamma}$, but not the association parameters $\boldsymbol{\alpha}$. To describe this pseudo-likelihood, suppose we let $f(y_{it}, r_{it}|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\gamma})$ denote the marginal distribution of (Y_{it}, R_{it}) at time t. We can write this distribution as

$$f(y_{it}, r_{it}|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) = f(y_{it}|\mathbf{x}_i, \boldsymbol{\beta}) f(r_{it}|y_{it}, \mathbf{x}_i, \boldsymbol{\gamma}),$$

where $f(y_{it}|\mathbf{x}_i, \boldsymbol{\beta})$ is given in (2.4), and $f(r_{it}|y_{it}, \mathbf{x}_i, \boldsymbol{\gamma})$ is Bernoulli, in which the probability of being observed is assumed to follow a logistic regression,

$$\pi_{it} = \pi_{it}(Y_{it}, \mathbf{x}_i, \boldsymbol{\gamma}) = pr(R_{it} = 1|y_{it}, \mathbf{x}_i, \boldsymbol{\gamma}) = \frac{\exp(\gamma_0 + \gamma_{1t} y_{it} + \boldsymbol{\gamma}_{2t}' \mathbf{x}_i)}{1 + \exp(\gamma_0 + \gamma_{1t} y_{it} + \boldsymbol{\gamma}_{2t}' \mathbf{x}_i)} . \tag{3.1}$$

If we only consider the data at time t, then our observed data likelihood would be

$$f(y_{it}, r_{it}|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\gamma})$$

if Y_{it} was observed, and would be

$$\sum_{y_{it}=0}^{1} f(y_{it}, r_{it}|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\gamma})$$

if Y_{it} was missing.

Then, the pseudo-likelihood of Troxel *et al.* (1998), which naively treats the observations at different times as independent, is

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\gamma}) &= \prod_{i=1}^{N}\prod_{t=1}^{T}[f(y_{it},r_{it}|\mathbf{x}_i,\boldsymbol{\beta},\boldsymbol{\gamma})]^{r_{it}}\left[\sum_{y_{it}=0}^{1} f(y_{it},r_{it}|\mathbf{x}_i,\boldsymbol{\beta},\boldsymbol{\gamma})\right]^{(1-r_{it})} \\
&= \prod_{i=1}^{N}\prod_{t=1}^{T}[f(y_{it}|\mathbf{x}_i,\boldsymbol{\beta})f(r_{it}|y_{it},\mathbf{x}_i,\boldsymbol{\gamma})]^{r_{it}} \\
&\quad \times\left[\sum_{y_{it}=0}^{1} f(y_{it}|\mathbf{x}_i,\boldsymbol{\beta})f(r_{it}|y_{it},\mathbf{x}_i,\boldsymbol{\gamma})\right]^{(1-r_{it})} \\
&= \prod_{i=1}^{N}\prod_{t=1}^{T}[f(y_{it}|\mathbf{x}_i,\boldsymbol{\beta})\pi_{it}]^{r_{it}}\left[\sum_{y_{it}=0}^{1} f(y_{it}|\mathbf{x}_i,\boldsymbol{\beta})(1-\pi_{it})\right]^{(1-r_{it})} \qquad (3.2)
\end{aligned}$$

This pseudo-likelihood is simply a product of terms at each measurement occasion: when an observation is present, the Bernoulli probability function $f(y_{it}|\mathbf{x}_i,\boldsymbol{\beta})$ is multiplied by the probability of being observed (π_{it}), and when the observation is missing, the product of $f(y_{it}|\mathbf{x}_i,\boldsymbol{\beta})$ and the missingness probability $(1-\pi_{it})$ is summed over the range of the missing measurement Y_{it}. Note that these marginal distributions are not a function of the association parameter $\boldsymbol{\alpha}$.

The maximum pseudo-likelihood estimate, $(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}})$, maximizes the log pseudo-likelihood. This estimate can be obtained by setting the first derivative of the log pseudo-likelihood, i.e., the pseudo-score vector,

$$S(\boldsymbol{\beta},\boldsymbol{\gamma}) = \frac{\partial}{\partial(\boldsymbol{\beta},\boldsymbol{\gamma})}\log\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\gamma}),$$

to $\mathbf{0}$, and solving for $(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}})$. Note that although the vector of observations on an individual are naively (and usually incorrectly) assumed to be independent, the resulting estimator of $(\boldsymbol{\beta},\boldsymbol{\gamma})$ is consistent. Heuristically, using method of moment ideas, $(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}})$ is consistent since it can be shown that $E[S(\boldsymbol{\beta},\boldsymbol{\gamma})] = \mathbf{0}$ at the true $(\boldsymbol{\beta},\boldsymbol{\gamma})$ if $f(y_{it},r_{it}|\mathbf{x}_i,\boldsymbol{\beta},\boldsymbol{\gamma})$ is correctly specified, and $(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}})$ is obtained as the solution to $S(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}}) = \mathbf{0}$. The maximum pseudo-likelihood estimate can be obtained using a Newton-Raphson algorithm, or the same EM-algorithm (Dempster *et al.*, 1977) that would be used if the (Y_{it},R_{it})'s are truly independent.

Thus, the naive pseudo-likelihood estimator $(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}})$ can be shown to be consistent and asymptotically normal. In this naive pseudo-likelihood approach, (Y_{it},R_{it}) is modeled without dependence on other measurement occasions. However, this 'marginal' model does not restrict the missingness at a given time to depend only upon the current, possibly unobserved response; rather, this marginal

approach requires only that the models for $f(y_{it}|\mathbf{x}_i, \boldsymbol{\beta})$ and $\text{pr}(R_{it} = 1|y_{it}, \mathbf{x}_i, \boldsymbol{\gamma})$ be correctly specified. As a result, the estimate $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$ could be biased if, for example, the logistic model for $\text{pr}(R_{it} = 1|y_{it}, \mathbf{x}_i, \boldsymbol{\gamma})$ in (3.1) is misspecified. Finally, we note that the negative second derivative of the log pseudo-likelihood will not provide a consistent estimator of the asymptotic variance; instead, the so-called "robust" or "sandwich" variance estimator can be used (White, 1982).

Overall, the naive pseudo-likelihood method is very appealing since only the marginal distributions of Y_{it} and R_{it} need to be correctly specified in order to obtain a consistent estimate of $\boldsymbol{\beta}$. That is, it is not necessary to specify the full joint distribution of $(\mathbf{Y}_i, \mathbf{R}_i)$ in order to obtain a consistent estimate of $\boldsymbol{\beta}$. The pseudo-likelihood approach is particularly attractive in this setting since the full likelihood can be far more complicated algebraically. In addition, ML estimation is computationally very demanding. Note, however, that an estimate of the association parameters $\boldsymbol{\alpha}$ cannot be obtained using this particular pseudo-likelihood approach. Furthermore, because the repeated measures over time are expected to be (positively) correlated, the maximum pseudo-likelihood estimate may in fact be very inefficient, as we have found in simulations in Section 6. In the following section, we describe a more efficient estimator of $\boldsymbol{\beta}$ based on a pseudo-likelihood approach that takes the correlation among repeated measures into account. The proposed pseudo-likelihood approach, based on all possible bivariate distributions, can provide more efficient estimators while still retaining much of the computational simplicity of the pseudo-likelihood approach of Troxel *et al.* (1998).

4. Pseudo-Likelihood Methods with Non-Ignorable Non-Monotone Missing Outcomes

In this section we propose a pseudo-likelihood approach, based on distributions at all pairs of times, for obtaining a more efficient estimate of $\boldsymbol{\beta}$, in addition to an estimate of the association parameter, $\boldsymbol{\alpha}$. We specify the joint distribution of $(Y_{is}, Y_{it}, R_{is}, R_{it}|\mathbf{x}_i)$ for each pair of times, but make no assumption about the joint distribution at three or more times. The pseudo-likelihood is based on the working assumption that $(Y_{is}, Y_{it}, R_{is}, R_{it}|\mathbf{x}_i)$ is independent of $(Y_{iu}, Y_{iv}, R_{iu}, R_{iv}|\mathbf{x}_i)$ for $(s, t) \neq (u, v)$.

First, we discuss the specification of the distribution of $(Y_{is}, Y_{it}, R_{is}, R_{it}|\mathbf{x}_i)$, which can be written as

$$f(y_{is}, y_{it}, r_{is}, r_{it}|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = f(y_{is}, y_{it}|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}) f(r_{is}, r_{it}|y_{is}, y_{it}, \mathbf{x}_i, \boldsymbol{\gamma}). \quad (4.1)$$

For ease of exposition, here, we assume that the pair (Y_{is}, Y_{it}) follow the

bivariate binary distribution due to Bahadur (1961),

$$\begin{aligned} f(y_{is}, y_{it}|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}) &= p_{is}^{y_{is}}(1-p_{is})^{(1-y_{is})} p_{it}^{y_{it}}(1-p_{it})^{(1-y_{it})} \\ &\quad \times \left\{ 1 + \rho_{ist} \frac{(y_{is}-p_{is})(y_{it}-p_{it})}{\sqrt{p_{is}(1-p_{is})p_{it}(1-p_{it})}} \right\}, \end{aligned} \tag{4.2}$$

where $\rho_{ist} = \rho_{ist}(\boldsymbol{\alpha}) = \text{Corr}(Y_{is}, Y_{it}|x_i)$. If one is interested in estimating the pairwise odds ratio instead of the correlation, then the bivariate Plackett (1965) distribution can be used. Other alternatives to (4.2), include the bivariate distributions proposed by Meester and MacKay (1994) and Heagerty (1999). However, for ease of exposition, we focus only on the bivariate Bahadur distribution here.

Note that similar assumptions can be made to specify $f(r_{is}, r_{it}|y_{is}, y_{it}, \mathbf{x}_i, \boldsymbol{\gamma})$. In particular, one can use the bivariate binary distribution

$$\begin{aligned} f(r_{is}, r_{it}|y_{is}, y_{it}, \mathbf{x}_i, \boldsymbol{\gamma}) &= \pi_{is}^{r_{is}}(1-\pi_{is})^{(1-r_{is})} \pi_{it}^{r_{it}}(1-\pi_{it})^{(1-r_{it})} \\ &\quad \times \left\{ 1 + \Gamma_{ist} \frac{(r_{is}-\pi_{is})(r_{it}-\pi_{it})}{\sqrt{\pi_{is}(1-\pi_{is})\pi_{it}(1-\pi_{it})}} \right\}, \end{aligned} \tag{4.3}$$

where $\Gamma_{ist} = \text{Corr}(R_{is}, R_{it}|y_{is}, y_{it}, \mathbf{x}_i)$ and $\pi_{it} = pr(R_{it} = 1|y_{it}, \mathbf{x}_i, \boldsymbol{\gamma})$ is given in (3.1).

In the bivariate binary distribution in (4.3), the probability of being observed ($R_{it} = 1$) at the current time can depend on the current (Y_{it}) and previous ($Y_{i,t-1}$) observations. However, this conditional probability does not have a logistic form, such as the model proposed by Diggle and Kenward (1994),

$$\begin{aligned} \eta_{it} &= pr(R_{it} = 1|r_{i,t-1}, y_{i,t-1}, y_{it}, \mathbf{x}_i, \boldsymbol{\gamma}) \\ &= \frac{\exp(\gamma_0 + \gamma_1 r_{i,t-1} + \gamma_2 y_{it} + \gamma_3 y_{i,t-1} + \boldsymbol{\gamma}_4' \mathbf{x}_i)}{1 + \exp(\gamma_0 + \gamma_1 r_{i,t-1} + \gamma_2 y_{it} + \gamma_3 y_{i,t-1} + \boldsymbol{\gamma}_4' \mathbf{x}_i)}. \end{aligned} \tag{4.4}$$

In particular, using (4.3), the conditional probability from the bivariate binary distribution is

$$\begin{aligned} &P(R_{it} = 1|r_{i,t-1}, y_{i,t-1}, y_{it}, \mathbf{x}_i) \\ &= \pi_{it} \left\{ 1 + \Gamma_{i,t-1,t} \frac{(r_{i,t-1}-\pi_{i,t-1})(1-\pi_{it})}{\sqrt{\pi_{i,t-1}(1-\pi_{i,t-1})\pi_{it}(1-\pi_{it})}} \right\}. \end{aligned} \tag{4.5}$$

In fact, the conditional probability that $R_{it} = 1$, given any R_{is} has form given by (4.5). Both (4.4) and (4.5) have simple forms.

Now, we discuss the proposed pseudo-likelihood, which is based on specifying the distribution of the data at all pairs of times on the same subject. First,

suppose there were only $T = 2$ time points. In this case, the pseudo-likelihood would equal the likelihood, which equals

$$\mathcal{L}_{12}(\boldsymbol{\Theta}) = L_1 \times L_2 \times L_3 \times L_4 \tag{4.6}$$

where

$$\begin{aligned}
L_1 &= \prod_{i=1}^{N} f(y_{i1}, y_{i2}, r_{i1}, r_{i2}|\mathbf{x_i}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})^{\mathbf{r_{i1}r_{i2}}} \\
L_2 &= \left[\sum_{y_{i1}} f(y_{i1}, y_{i2}, r_{i1}, r_{i2}|\mathbf{x_i}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})\right]^{(1-r_{i1})r_{i2}} \\
L_3 &= \left[\sum_{y_{i2}} f(y_{i1}, y_{i2}, r_{i1}, r_{i2}|\mathbf{x_i}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})\right]^{r_{i1}(1-r_{i2})} \\
L_4 &= \left[\sum_{y_{i1}, y_{i2}} f(y_{i1}, y_{i2}, r_{i1}, r_{i2}|\mathbf{x_i}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})\right]^{(1-r_{i1})(1-r_{i2})} .
\end{aligned}$$

To be more explicit about (4.6), if a subject is observed at both times 1 and 2, i.e., $(r_{i1} = 1, r_{i2} = 1)$, then that subject's contribution to the likelihood is

$$f(y_{i1}, y_{i2}, r_{i1}, r_{i2}|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = f(y_{i1}, y_{i2}, r_{i1}, r_{i2}|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})^{r_{i1}r_{i2}} \ ;$$

if a subject is missing at time 1 and observed at time 2, i.e., $(r_{i1} = 0, r_{i2} = 1)$, then that subject contributes the following marginal distribution (summed over possible values of y_{i1}) to the likelihood

$$\sum_{y_{i1}} f(y_{i1}, y_{i2}, r_{i1}, r_{i2}|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \left[\sum_{y_{i1}} f(y_{i1}, y_{i2}, r_{i1}, r_{i2}|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})\right]^{(1-r_{i1})r_{i2}} ;$$

if a subject is missing at time 2 and observed at time 1, i.e., $(r_{i1} = 1, r_{i2} = 0)$, then that subject contributes the following marginal distribution (summed over possible values of y_{i2}) to the likelihood

$$\sum_{y_{i2}} f(y_{i1}, y_{i2}, r_{i1}, r_{i2}|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \left[\sum_{y_{i2}} f(y_{i1}, y_{i2}, r_{i1}, r_{i2}|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})\right]^{r_{i1}(1-r_{i2})} ;$$

finally, if a subject is missing at both times 1 and 2, i.e., $(r_{i1} = 0, r_{i2} = 0)$, then that subject contributes the following marginal distribution (summed over possible values of both y_{i1} and y_{i2}) to the likelihood

$$\sum_{y_{i1}, y_{i2}} f(y_{i1}, y_{i2}, r_{i1}, r_{i2}|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \left[\sum_{y_{i1}, y_{i2}} f(y_{i1}, y_{i2}, r_{i1}, r_{i2}|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})\right]^{(1-r_{i1})(1-r_{i2})} .$$

In the general pseudo-likelihood ($T > 2$), we naively assume that $(Y_{ij}, Y_{ik}, R_{ij}, R_{ik})$ and $(Y_{i\ell}, Y_{im}, R_{i\ell}, R_{im})$ are independent. With T time points, there are $T(T-1)/2$ pairs of times, and the pseudo-likelihood will be a product of $T(T-1)/2$ terms, with the term corresponding to times s and t being identical to (4.6), after replacing times 1 and 2 in (4.6) with times s and t. In particular, naively assuming that $(Y_{ij}, Y_{ik}, R_{ij}, R_{ik})$ and $(Y_{i\ell}, Y_{im}, R_{i\ell}, R_{im})$ are independent, the proposed pseudo-likelihood for $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})'$ is given by the product of the likelihood in (4.6) over all pairs of times,

$$\mathcal{L}_2(\boldsymbol{\Theta}) = L_5 \times L_6 \times L_7 \times L_8 = \prod_{i=1}^{N} \prod_{s<t} L_{ist}(\boldsymbol{\Theta}), \tag{4.7}$$

where

$$\begin{aligned}
L_5 &= \prod_{i=1}^{N} \prod_{s<t} f(y_{is}, y_{it}, r_{is}, r_{it} | \mathbf{x_i}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})^{\mathbf{r_{is} r_{it}}} \\
L_6 &= \left[\sum_{y_{is}} f(y_{is}, y_{it}, r_{is}, r_{it} | \mathbf{x_i}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \right]^{(1-r_{is}) r_{it}} \\
L_7 &= \left[\sum_{y_{it}} f(y_{is}, y_{it}, r_{is}, r_{it} | \mathbf{x_i}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \right]^{r_{is}(1-r_{it})} \\
L_8 &= \left[\sum_{y_{is}, y_{it}} f(y_{is}, y_{it}, r_{is}, r_{it} | \mathbf{x_i}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \right]^{(1-r_{is})(1-r_{it})}
\end{aligned}$$

and L_{ist} is the contribution from $(Y_{is}, Y_{it}, R_{is}, R_{it})$. The pseudo-score is

$$S_2(\boldsymbol{\Theta}) = \sum_{i=1}^{N} S_{2i}(\boldsymbol{\Theta}) = \sum_{i=1}^{N} \sum_{s<t} \frac{\partial}{\partial \boldsymbol{\Theta}} \log[L_{ist}(\boldsymbol{\Theta})], \tag{4.8}$$

and the maximum pseudo-likelihood estimate, $\widehat{\boldsymbol{\Theta}}$, is the solution to $S_2(\widehat{\boldsymbol{\Theta}}) = \mathbf{0}$. Heuristically, using method of moment ideas, assuming that $f(y_{is}, y_{it}, r_{is}, r_{it} | \mathbf{x_i}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ is correctly specified, $\widehat{\boldsymbol{\Theta}}$ is consistent since $E[S_2(\boldsymbol{\Theta})] = \mathbf{0}$ at the true $\boldsymbol{\Theta}$ if the bivariate distributions are correctly specified, and $\widehat{\boldsymbol{\Theta}}$ is obtained as the solution to $S_2(\widehat{\boldsymbol{\Theta}}) = \mathbf{0}$. The maximum pseudo-likelihood estimate can be obtained using a Newton-Raphson algorithm, or the same EM-algorithm (Dempster *et al.*, 1977) that would be used if $(Y_{ij}, Y_{ik}, R_{ij}, R_{ik})$ and $(Y_{i\ell}, Y_{im}, R_{i\ell}, R_{im})$ are truly independent. The estimate $\widehat{\boldsymbol{\Theta}}$ is also asymptotically multivariate normal. However, a robust estimator of variance is required, since $(Y_{ij}, Y_{ik}, R_{ij}, R_{ik})$ and $(Y_{i\ell}, Y_{im}, R_{i\ell}, R_{im})$ will, in general, be correlated. The appropriate adjustment

takes the form of a sandwich estimator commonly used in statistical practice. In particular,

$$n^{\frac{1}{2}}(\hat{\mathbf{\Theta}} - \mathbf{\Theta}) \quad \xrightarrow{\mathcal{L}} \quad \mathcal{N}(0, \Sigma),$$

where

$$\Sigma \quad = \quad \left[\frac{1}{n} E\left\{\frac{\partial S_2(\mathbf{\Theta})}{\partial \mathbf{\Theta}}\right\}\right]^{-1} \frac{1}{n}\sum_{i=1}^{n} E\left\{S_{2i}(\mathbf{\Theta}) S_{2i}'(\mathbf{\Theta})\right\} \left[\frac{1}{n} E\left\{\frac{\partial S_2(\mathbf{\Theta})}{\partial \psi}\right\}\right]^{-1} \quad (4.9)$$

The variance estimate is obtained by replacing $\mathbf{\Theta}$ with $\hat{\mathbf{\Theta}}$ in (4.9) to get $\hat{\Sigma}$.

We note that the pseudo-likelihood in (4.7) can be considered an extension to non-ignorable missingness of the pseudo-likelihood proposed by Le Cessie and Van Houwelingen (1994). Le Cessie and Van Houwelingen (1994) proposed a pseudo-likelihood for repeated binary data in which missing observations must be missing completely at random. Briefly, Le Cessie and Van Houwelingen's pseudo-likelihood is similar to (4.7), except without the part that contains the missing data mechanism, i.e., Le Cessie and Van Houwelingen's pseudo-likelihood is

$$\prod_{i=1}^{N}\prod_{s<t} f(y_{is}, y_{it}|\mathbf{x_i}, \boldsymbol{\beta}, \boldsymbol{\alpha})^{\mathbf{r_{is}r_{it}}} \left[\sum_{\mathbf{y_{is}}} \mathbf{f}(\mathbf{y_{is}}, \mathbf{y_{it}}|\mathbf{x_i}, \boldsymbol{\beta}, \boldsymbol{\alpha})\right]^{\mathbf{(1-r_{is})r_{it}}}$$
$$\times \left[\sum_{y_{it}} f(y_{is}, y_{it}|\mathbf{x_i}, \boldsymbol{\beta}, \boldsymbol{\alpha})\right]^{r_{is}(1-r_{it})}.$$

We also note that, if one specifies $f(y_{is}, y_{it}|\mathbf{x_i}, \boldsymbol{\beta}, \boldsymbol{\alpha})$ as in (4.2), but specifies the missingness model as in (4.4), to determine the bivariate density $f(y_{is}, y_{it}, r_{is}, r_{it}|\mathbf{x_i}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ would first require specification of the full density $f_{y,r}(\mathbf{y_i}, \mathbf{r_i}|\mathbf{x_i}, \boldsymbol{\theta}, \boldsymbol{\gamma})$; then, one needs to sum over the appropriate times to get the bivariate density. Unfortunately, since one needs to specify the full density over all times in this case, the bivariate pseudo-likelihood (4.7) will not have the robust property that only the bivariate distributions need to be correctly specified. Further, one could consider forming trivariate or multivariate pseudo-likelihoods; however, the computational complexity increases, and we also loose the robustness property that only the first two moments need to be correctly specified. Finally, even though the pseudo-likelihood is less computationally intensive than maximum likelihood, it could be much more computationally intensive than Troxel *et al.*'s (1998) pseudo-likelihood that is formed by naively assuming that the longitudinal binary measurements are independent over time. In Troxel et al.'s (1998) pseudo-likelihood, we need to form T 'observations' for each subject; for ours, we need to form

$T(T-1)/2$. For most of the studies we consider, in which T is small, say less than 20, the difference in computing is negligible. However, when T is large, such as in a diary study in which the outcome is to be measured every day for a year, then the difference in computational time could be substantial.

5. Six Cities Example

We now present an analysis of the Six Cities Study described in the introduction. The Six Cities study is a longitudinal study of the health effects of air pollution (Ware *et al.*, 1984). The dataset contains records on $n = 3331$ children who resided in Kingston-Harriman, Tennessee or Portage, Wisconsin (two of the participating cities). The response of interest at time t ($t = 1, 2, 3, 4$) is the wheezing of the child, $Y_{it} = 1$ if wheeze, 0 if no wheeze. As discussed in the introduction, wheeze is missing between 30% and 40% at each of the four times, and the missing data patterns are non-monotone. Data is also collected on the covariates maternal smoking (measured in cigarettes per day) at baseline (age 9), denoted by smoke_i, and city, denoted by city_i, which equals 1 if Kingston-Harriman or 0 if Portage. Children of mothers who smoke at baseline and who reside in the more polluted city (Kingston-Harriman) are expected to have higher rates of wheeze.

Thus, we model the probability of wheeze at a given age as a function of the Child's age, maternal smoking at baseline and the city of residence,

$$\ln\left[\frac{p_{it}}{1-p_{it}}\right] = \beta_0+\beta_1\text{city}_i+\beta_2\text{smoke}_i+\beta_3 I(t=2)+\beta_4 I(t=3)+\beta_5 I(t=4), \quad (5.1)$$

for $t = 1, 2, 3, 4$ and where $I(\cdot)$ are indicator variables. Finally, when using our proposed pseudo-likelihood, we assumed an unstructured correlation between the responses at times s and t, so that $Corr(Y_{is}, Y_{it}|\mathbf{x}_i) = \rho_{st}$.

It is reasonable to conjecture that wheezing is nonignorably missing since a child may not come in for a doctor's visit if he or she is not wheezing, so that there is no need to make the visit. Thus, it is plausible to propose a nonignorable missing data mechanism for wheezing. Thus, we fit the following missingness model that included the current outcome,

$$\ln\left[\frac{\pi_{it}}{1-\pi_{it}}\right] = \gamma_0+\gamma_1 y_{it}+\gamma_2\text{city}_i+\gamma_3\text{smoke}_i+\gamma_4 I(t=2)+\gamma_5 I(t=3)+\gamma_6 I(t=4)\ . \quad (5.2)$$

Finally, when using our proposed pseudo-likelihood, we assumed an unstructured correlation between the missingness indicators at times s and t, so that

$$Corr(R_{is}, R_{it}|y_{is}, y_{it}, \mathbf{x}_i) = \Gamma_{st}\ . \quad (5.3)$$

We also fit our pseudo-likelihood assuming $\Gamma_{st} = 0$.

Table 2 gives estimates and standard errors for the parameters $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ for all the methods fit, and Table 3 gives the estimates of the missingness parameters, $\boldsymbol{\gamma}$. The methods include Troxel *et al.* (1998), denoted by PSL1; our pseudo-likelihood in which we estimated the missingness correlations Γ_{st}, which we denote by PSL2CORR; and our pseudo-likelihood in which we set $\Gamma_{st} = 0$, which we denote by PSL2. Further, we fit Troxel *et al.*'s (1998) pseudo-likelihood under the assumption of missing completely at random; in this case, Troxel *et al.*'s estimate reduces to the independence generalized estimating equations of Liang and Zeger (1986). We denote this estimate IGEE. Finally, we also fit our pseudo-likelihood under the assumption of missing completely at random, which, as discussed earlier, is equivalent to the pseudo-likelihood proposed by Le Cessie and Van Houwelingen (1994); we denote this estimate by PSL2MCAR.

From Table 2, we see that the parameter estimates of $\boldsymbol{\beta}$ and estimated standard errors from Troxel *et al.*'s PSL1 are very different from the estimates from the other approaches. In particular, the city by smoking interaction is at least three time larger for PSL1 than the estimate obtained by any other method, and the estimate is significant using PSL1, whereas it is not significant using any other approach. Further, the time 4 estimate is about 1/3 the magnitude using PSL1 compared to any other method, and is non-significant. The reason for these differences lies in the estimate of the missingness model in Table 3, in which PSL1 estimates a significant non-ignorable effect of wheeze at the current time, whereas the other approaches do not. In particular, using PSL1, subjects who wheeze are estimated to have $\exp(-4.120) = 0.016$ the odds of being observed compared to subjects who do not wheeze. Further, we also see that there are differences between PSL2CORR, in which we estimate the missingness correlations, and PSL2, in which we assume these correlations are 0. In particular, the times 2 and 3 effects in the model for p_{it} are significant using PSL2CORR, but not using PSL2. The estimates of $\boldsymbol{\beta}$ are not much different using PSL2CORR and PSL2, but the estimated standard errors appear much larger using PSL2. The estimated standard errors for the missingness parameters are also much larger for PSL2 compared to PSL2CORR. Thus, taking the possible correlation of the missingness indicators into account appears to reduce the variance of the estimated parameters. Looking at the estimated correlations $\widehat{\boldsymbol{\Gamma}}_{st}$ from PSL2CORR, we see from table 4 that the missingness indicators 1 year apart in time are highly correlated (correlation greater than .4), the missingness indicators 2 years apart in time are mildly correlated (correlation about .16), and missingness indicators 3 years apart in time have correlation close to 0. The estimates assuming MCAR from IGEE and PSL2MCAR were similar to each other, and similar to PSL2CORR.

We also performed sensitivity analyses for PSL1, PSL2CORR, and PSL2. In particular, we fit various non-ignorable missingness mechanisms by dropping different combinations of city, smoking, and time from the missingness model. We note that, regardless of what combination of city, smoking, and time we have in the missingness model, if we drop the outcome wheeze Y_{it} from the missingness model, we get the IGEE estimates using PSL1 and the PSL2MCAR estimates using either PSL2CORR or PSL2. In the sensitivity analysis, we found that the estimates of $\boldsymbol{\beta}$ changed very little from that given in Table 2 for any non-ignorable model for PSL2CORR and PSL2. However, for PSL1, the estimate of $\boldsymbol{\beta}$ changed a lot depending on the non-ignorable missingness model. For example, if, for PSL1, we let the non-ignorable missingness model only depend on wheeze, but not city, smoking, or time, we found that both the Newton-Raphson algorithm and the EM-algorithm did not converge (the Newton-Raphson algorithm stopped when the estimated coefficient of wheeze in the missingness model equaled -13). This example illustrates that different non-ignorable models and different methods to estimate the parameters can lead to different, and possibly conflicting estimates. To look at the properties of these approaches further, we performed simulations in the following section.

6. Simulation Study

In this section we present simulations to study the finite sample performance of the estimators discussed in the previous section: PSL1 (Troxel *et al.*, 1998); PSL2CORR (our pseudo-likelihood in which we estimate the missingness correlations Γ_{st}); PSL2 (our pseudo-likelihood with $\Gamma_{st} = 0$); IGEE (independence generalized estimating equations); and PSL2MCAR (our pseudo-likelihood under MCAR, Le Cessie and Van Houwelingen, 1994). In the simulations, we let $T = 3$, so that the full likelihood is tractable, and we can also compute the MLE for the simulations.

Thus, for simplicity, in the simulation study, we consider the case of a *trivariate* binary response and a simple two group configuration, e.g. active treatment versus placebo. Subjects are assumed to belong to either group with equal probability. We denote this dichotomous covariate indicating group membership by x_i, which equals 1 if group 1, 0 if group 0. To specify the true underlying joint distribution of the binary responses, (Y_{i1}, Y_{i2}, Y_{i3}), we choose the model for correlated binary data first described by Bahadur (1961), and later by Cox (1972). In general, with binary responses at each of T times, the joint distribution of an individual's responses at the T times is multinomial with 2^T probabilities corresponding to the 2^T possible response profiles. Thus, in Bahadur's correlated binary model, the joint distribution of an individual's responses at the three times is multinomial with $2^3 = 8$ cell probabilities, $pr(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, Y_{i3} = y_{i3}|x_i)$,

where $y_{it} = 0, 1$. If we define the standardized variable Z_{it} to be

$$Z_{it} = \frac{Y_{it} - p_{it}}{\sqrt{p_{it}(1-p_{it})}} ,$$

then the Bahadur model of the 2^3 multinomial probabilities is

$$\begin{aligned} \mathrm{pr}(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, Y_{i3} = y_{i3} | x_i, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \left\{ \prod_{t=1}^{3} p_{it}^{y_{it}} (1-p_{it})^{(1-y_{it})} \right\} \\ \times \left\{ 1 + \rho_{12} z_{i1} z_{i2} + \rho_{13} z_{i1} z_{i3} + \rho_{23} z_{i2} z_{i3} + \rho_{123} z_{i1} z_{i2} z_{i3} \right\} , \end{aligned} \tag{6.1}$$

where

$$\rho_{st} = Corr(Y_{is}, Y_{it}) = \frac{E[(Y_{is} - p_{is})(Y_{it} - p_{it}) | x_i]}{\sqrt{p_{is}(1-p_{is}) p_{it}(1-p_{it})}};$$

$$\rho_{123} = \frac{E[(Y_{i1} - p_{i1})(Y_{i2} - p_{i2})(Y_{i3} - p_{i3}) | x_i]}{\sqrt{p_{i1}(1-p_{i1}) p_{i2}(1-p_{i2}) p_{i3}(1-p_{i3})}};$$

and

$$p_{it} = \frac{\exp[\beta_0 + \beta_x x_i + \beta_t (t-1)]}{1 + \exp[\beta_0 + \beta_x x_i + \beta_t (t-1)]}$$

for t=1,2,3. The parameter ρ_{123} can be thought of as a "three-way" association parameter, and $\boldsymbol{\alpha} = [\rho_{12}, \rho_{13}, \rho_{23}, \rho_{123}]'$.

For the simulation study, the parameters of the true model are as follows. The marginal regression parameters are $[\beta_0, \beta_x, \beta_t] = [-0.25, 0.5, 0.20]$. A variety of different correlation structures were examined and the same overall pattern of results were obtained. For simplicity, we present the results from an unstructured correlation, $[\rho_{12}, \rho_{13}, \rho_{23}] = [.4, .3, .5]$. The three-way correlation, ρ_{123}, is held fixed at zero.

We performed separate simulations with three true non-ignorable missingness mechanisms. First, we let the missingness indicators be independent at the three times, with

$$\pi_{it} = pr(R_{it} = 1 | y_{it}, x_i, \boldsymbol{\gamma}) = \frac{\exp[\gamma_0 + \gamma_1 x_i + \gamma_2 (t-1) + \gamma_3 y_{it}]}{1 + \exp[\gamma_0 + \gamma_1 x_i + \gamma_2 (t-1) + \gamma_3 y_{it}]}. \tag{6.2}$$

For the simulation study, the parameters of the true model in (6.2) are

$$\gamma_0 = -.5;\ \gamma_1 = 1;\ \gamma_2 = .2;\ \gamma_3 = 1\ . \tag{6.3}$$

Here, missingness at a given time depends upon group membership, time, and the possibly missing outcome at that time. In this mechanism, non-monotone

missingness can occur in that an outcome can be missing at time s ($R_{is} = 0$), but observed at a future time t ($R_{it} = 1$ for $t > s$). Given the parameters in (6.3), the percentage missing at a given time is between 30% and 40%. Then, the full distribution $f_r(\mathbf{r}_i|\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\gamma})$ is

$$\text{pr}[R_{i1} = r_{i1}, R_{i2} = r_{i2}, R_{i3} = r_{i3}|y_{i1}, y_{i2}, y_{i3}, x_i, \boldsymbol{\gamma}] = \prod_{t=1}^{3} \pi_{it}^{r_{it}}(1 - \pi_{it})^{(1-r_{it})}. \quad (6.4)$$

For the second non-ignorable missingness mechanism, we let $f_r(\mathbf{r}_i|\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\gamma})$ follow a Bahadur distribution, similar to that of $\mathbf{Y}_i$. If define the standardized variable U_{it} to be

$$U_{it} = \frac{R_{it} - \pi_{it}}{\sqrt{\pi_{it}(1 - \pi_{it})}},$$

then the Bahadur distribution is

$$\begin{aligned}\text{pr}\{R_{i1} = r_{i1}, R_{i2} = r_{i2}, R_{i3} = r_{i3}|y_{i1}, y_{i2}, y_{i3}, x_i, \boldsymbol{\gamma}\} &= \left\{\prod_{t=1}^{3} \pi_{it}^{r_{it}}(1 - \pi_{it})^{(1-r_{it})}\right\} \\ &\times \left\{1 + \Gamma_{12}u_{i1}z_{i2} + \Gamma_{13}u_{i1}u_{i3} + \Gamma_{23}u_{i2}u_{i3} + \Gamma_{123}u_{i1}u_{i2}u_{i3}\right\}, \quad (6.5)\end{aligned}$$

where

$$\Gamma_{st} = Corr(R_{is}, R_{it}|y_{i1}, y_{i2}, y_{i3}, x_i)$$

$$\Gamma_{123} = \frac{E[(R_{i1} - \pi_{i1})(R_{i2} - \pi_{i2})(R_{i3} - \pi_{i3})|y_{i1}, y_{i2}, y_{i3}, x_i]}{\sqrt{\pi_{i1}(1 - \pi_{i1})\pi_{i2}(1 - \pi_{i2})\pi_{i3}(1 - \pi_{i3})}};$$

and π_{it} is given in (6.2). In the simulations, we set the true parameters for the model for π_{it} equal to those given in (6.3), and we set $(\Gamma_{12}, \Gamma_{13}, \Gamma_{23}, \Gamma_{123}) = (.2, .1, .3, 0)$. Note, the missingness model in (6.4) is a special case of (6.5) in which we set $(\Gamma_{12}, \Gamma_{13}, \Gamma_{23}, \Gamma_{123}) = (0, 0, 0, 0)$.

The third missingness mechanism will allow the probability $R_{it} = 1$ to depend on the values of x_i, Y_{it}, $Y_{i,t-1}$, and $R_{i,t-1}$, through a logistic regression,

$$\begin{aligned}\eta_{it} &= pr(R_{it} = 1|r_{i,t-1}, \mathbf{y_i}, \mathbf{x_i}, \boldsymbol{\gamma})] \\ &= \frac{\exp\{\gamma_0 + \gamma_1 x_i + \gamma_2 t + \gamma_3 y_{it} + I[t > 1](\gamma_4 y_{i,t-1} + \gamma_5 r_{i,t-1})\}}{1 + \exp\{[\gamma_0 + \gamma_1 x_i + \gamma_2 t + \gamma_3 y_{it} + I[t > 1](\gamma_4 y_{i,t-1} + \gamma_5 r_{i,t-1})\}} \quad (6.6)\end{aligned}$$

We see in (6.6), for time 1, $(\gamma_4 y_{i,t-1} + \gamma_5 r_{i,t-1})$ is dropped from the model. Then, the full distribution $f_r(\mathbf{r}_i|\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\gamma})$ is

$$\text{pr}[R_{i1} = r_{i1}, R_{i2} = r_{i2}, R_{i3} = r_{i3}|y_{i1}, y_{i2}, y_{i3}, x_i, \boldsymbol{\gamma}] = \prod_{t=1}^{3} \eta_{it}^{r_{it}}(1 - \eta_{it})^{(1-r_{it})}. \quad (6.7)$$

The missingness model in (6.2) can be considered a special case of (6.6) when $\gamma_4 = \gamma_5 = 0$. If (6.7) is the correct missingness mechanism, then all of the pseudo-likelihoods proposed (both Troxel et. al.'s and ours) will be misspecified. We performed simulations with the missingness mechanism in (6.7) to see how much bias arises in the pseudo-likelihoods when the missingness mechanism is misspecified. We performed 1000 simulations for each configuration, with a sample size of $n = 500$. We chose a sample size of 500 because of the number of parameters $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ to estimate.

In the simulation in Table 4, the missingness mechanism is given by (6.4), corresponding to independent missingness indicators at the three times. For PSL1, PSL2CORR, PSL2, and maximum likelihood estimation (MLE), the missingness mechanism is correctly specified. We see that, despite the correct specification, PSL1 is slightly negatively biased for β_x and β_t, and the coverage probability for PSL1 is poor, due to the slight bias in the estimate, and the poor performance of the sandwich variance estimate. In simulations not shown, for a sample size of $n = 1000$, this slight bias goes away, and the bias in the sandwich variance estimate decreased. PSL2CORR, PSL2, and MLE are all approximately unbiased. We see that our proposed pseudo-likelihood estimates, PSL2CORR and PSL2, are both slightly more efficient than the asymptotically efficient MLE. Further, the coverage probability of both PSL2CORR and PSL2 are at the nominal level of 95%. Looking at the simulation variance, and neglecting the small bias for PSL1 (or, equivalently, using the simulation mean square error), we see that PSL2CORR and PSL2 are much more efficient than PSL1. The estimates assuming MCAR, IGEE and PSL2MCAR, are highly biased.

In the simulation in Table 5, the missingness mechanism is given by (6.5), corresponding to correlated missingness indicators at the three times. For PSL1, PSL2CORR, and maximum likelihood estimation (MLE), the missingness mechanism is correctly specified. For PSL2, the missingness mechanism is incorrectly specified. We see that again, despite the correct specification, PSL1 is slightly negatively biased for β_x and β_t, and the coverage probability for PSL1 is poor, due again to the slight bias in the estimate, and the poor performance of the sandwich variance estimate. PSL2 also give biased estimates, due to the fact that the missingness mechanism is incorrectly specified. PSL2CORR and MLE are both approximately unbiased. Looking at the simulation variance, we see that our proposed pseudo-likelihood estimate PSL2CORR is slightly less efficient than the asymptotically efficient MLE. Further, the coverage probability of PSL2CORR is at the nominal level of 95%. The estimates assuming MCAR, IGEE and PSL2MCAR, are again highly biased.

In the simulation in Table 6, the missingness mechanism is given by (6.7), corresponding to logistic missingness models given current and previous outcomes.

We did not fit maximum likelihood for this approach since we were not interested in efficiency because we know that the pseudo-likelihood estimates will be biased. PSL1 and PSL2CORR have the least bias, whereas all others, including PSL2, have high bias. This simulation, as well as the simulation in Table 5, suggest that PSL2 is not very robust to mis-specification of the missingness model, whereas PSL2CORR, due to estimation of correlations between the missingness indicators, is more robust. These simulations suggest that, in general, PSL2CORR is the method of choice.

7. Discussion

We have proposed pseudo-likelihoods for the estimation of marginal models for longitudinal binary data with non-monotone non-ignorable missing outcomes. Unlike the full likelihood, the pseudo-likelihoods require specification of the distribution for the data at all pairs of times on the same subject. Further, compared to maximum likelihood, which requires the full likelihood to be correctly specified in order to obtain consistent estimates, the pseudo- likelihood estimates are consistent as long as the bivariate distributions are correctly specified. Because of the broad range of possible missing data configurations and underlying probability distributions generating the data, it is difficult to draw definitive conclusions from simulation studies. We can only make general suggestions. Based on our simulation studies, we have examined the efficiency and bias of our proposed pseudo-likelihood estimates, and found that our pseudo-likelihood approach PSL2CORR, in which we estimate the missingness correlations Γ_{st}, has the best properties. In particular, when the missingness mechanism is correctly specified, PSL2CORR is either more efficient, or almost as efficient, as the asymptotic efficient MLE. Further, when the missingness mechanism is misspecified, PSL2CORR appears to have small bias. Further, PSL2CORR is much more efficient than the pseudo-likelihood proposed by Troxel *et al.* (1998).

For non-ignorable missingness models, one can encounter multimodal likelihood or pseudo-likelihood surfaces: the likelihood consists of gentle peaks and valleys near the solution. With both actual and simulated data, for our pseudo-likelihood PSL2CORR, using different starting points, the Newton-Raphson algorithm always converged to the same maximum. This was not true for Troxel *et al.*'s PSL1, although we were always able to find a unique maximum. Apparently, specifying and estimating correlations for the repeated binary outcomes and the missingness mechanism, led to more 'information', and a more concave pseudo-likelihood using PSL2CORR. Further, in general, it is well- known that estimates from non-ignorable models are quite sensitive to modelling assumptions. We view the pseudo-likelihood estimators proposed here as another tool for conducting sensitivity analysis for non-ignorable models, which can be used

to explore departures from ignorable missingness models.

Acknowledgments

The authors are grateful for the support provided by the following grants from the United States' Institutes of Health: HL 69800, AHRQ 10871, HL52329, HL61769, CA 70101, and CA 74015.

References

Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. In *Studies in Item Analysis and Prediction*, (Edited by H. Solomon, 158-68. Stanford Mathematical Studies in the Social Sciences VI. Stanford University Press.

Baker, S. G. (1995). Marginal regression for repeated binary data with outcomes subject to nonignorable nonresponse. *Biometrics*, **51**, 1042-1052.

Baker, S. G., and Laird, N. M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association* **83**, 62-69.

Cox, D. R. (1972). The analysis of multivariate binary data. *Applied Statistics* **21**, 113-20.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1-38.

Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal analysis (with discussion). *Applied Statistics* **43**, 49-93.

Fitzmaurice G., Molenberghs, G. and Lipsitz, S. R. (1996). Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royall Statist. Soc. Ser. B* **57**, 691-704.

Gong, G. and Samaniego, F. (1982). Pseudo maximum likelihood estimation: Theory and applications. *Ann. Statist.* **9**, 861-869.

Heagerty P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics* **55**, 688-698.

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

Le Cessie, S. and Van Houwelingen J. C. (1994). Logistic regression for correlated binary data. *Applied Statistics* **43**, 95-108.

Liang, K. Y. and Self, S. G. (1996). On the asymptotic behavior of the pseudolikelihood ratio test statistic. *Journal of the Royall Statist. Soc. Ser. B* **58**, 785-796.

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

Little, R. J. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. Wiley & Sons.

Meester S. G. and MacKay J. (1994). A parametric model for cluster correlated categorical data. *Biometrics.* **50**, 954-963.

Molenberghs, G. and Lesaffre, E. (1994). Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. *J. of the Amererican Statist. Assoc.* **89**, 633-644.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimisation. *Computer Journal* **7**, 303-313.

Troxel, A. B., Lipsitz, S. R., and Harrington, D. P. (1998). Marginal models for the analysis of longitudinal measurements subject to nonignorable non-monotone missing data. Submitted to *Biometrika*.

Ware, J. H., Dockery, D. W., Spiro, A. III, Speizer, F. E. and Ferris, B. G. Jr. (1984). Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases* **129**, 366-374.

Received July 19, 2005; accepted January 17, 2006.

Michael Parzen
Goizueta School of Business
Emory University

Stuart R. Lipsitz
Department of Biometry and Epidemiology
Medical University of South Carolina

Garrett M. Fitzmaurice
Department of Biostatistics
Harvard School of Public Health,

Joseph G. Ibrahim
Department of Biostatistics
The University of North Carolina

Andrea Troxel
Department of Biostatistics
Columbia University School of Public Health,

Geert Molenberghs
Center for Statistics
Limburgs Universitair