Statistica Sinica **22** (2012), 1103-1121
doi:http://dx.doi.org/10.5705/ss.2010.029

SEMIPARAMETRIC PSEUDO LIKELIHOODS FOR LONGITUDINAL DATA WITH OUTCOME-DEPENDENT NONMONOTONE NONRESPONSE

Deyuan Jiang and Jun Shao

University of Wisconsin and East China Normal University

Abstract: In a study with longitudinal outcomes, the outcome nonresponse mechanism often depends on the observed or unobserved value of the outcome. When nonresponse is monotone in the sense that a subject having a missing outcome at time t is not observed after time t, Tang, Little, and Raghunathan (2003) developed a semiparametric pseudo-likelihood method for the estimation of parameters of interest. In practice, however, nonresponse is often not monotone and a direct application of their method discards observed data from subjects having nonmonotone nonresponse, which may result in inefficient estimators. We extend the idea in Tang, Little, and Raghunathan (2003) to nonmonotone nonresponse and construct a semiparametric pseudo-likelihood that utilizes all observed data. Asymptotic normality of the maximum pseudo-likelihood estimators is established. An application is made to the household income data from the Health and Retirement Study. Simulation results are also presented to examine finite sample properties of the proposed estimators.

Key words and phrases: Efficiency, nonignorable missing, semiparametric likelihood, sequential estimation.

1. Introduction

Longitudinal data at multiple time points are often collected from a sampled subject in medical, health, economical, and social studies. Let Y_t denote the outcome at time point t, $Y = (Y_1, \ldots, Y_T)$, and X be a time-independent covariate vector or a time-dependent covariate $X = (X_1, \ldots, X_T)$ with each X_t being a covariate vector. In the Health and Retirement Study (HRS) discussed in Section 4, households of seniors in the United States are surveyed biannually with Y_t being the household income and X being household characteristics such as the years of education, years of working experience, and health status. The statistical analysis typically aims to estimate or make inference on some unknown parameters in $p(Y|X)$ or $p(Y)$, where $p(A|B)$ or $p(A)$ is a generic notation for the probability density of A conditional on B or the marginal probability density of A.

In many studies Y has nonresponse although values of X are all observed. Nonresponse is monotone if Y_{t+1} is missing whenever Y_t is missing, $t = 1, \ldots, T$, which is also referred to as dropout. Nonresponse is often nonmonotone. In the HRS, for example, a household not responding in a particular year may respond to the next biannual survey.

Let $R_t = 1$ if Y_t is observed and $R_t = 0$ if Y_t is missing, $t = 1, \ldots, T$, and let $R = (R_1, \ldots, R_T)$. In the presence of nonresponse, our analysis has to focus on $p(Y, R|X) = p(R|Y, X)p(Y|X)$. If nonresponse is ignorable in the sense that $p(R|Y, X) = p(R|Y_o, X)$, where Y_o is the observed part of Y, then parameters in $p(Y|X)$ can be estimated without requiring any further assumption on $p(R|Y_o, X)$ (see, e.g., Little and Rubin (2002)). For longitudinal Y, the ignorable nonresponse assumption may be reasonable when nonresponse is monotone, but it is not natural when nonresponse is nonmonotone. In the HRS, for example, it is hard to imagine how a household's response status on the income in a particular year is related to its past observed incomes but not past unobserved incomes, and is related to its future "observed" income when we do not know whether that future income is missing or not. When nonresponse is nonignorable, i.e., $p(R|Y, X)$ depends on observed and unobserved components of Y, we must impose a parametric model on either $p(R|Y, X)$ or $p(Y|X)$. Otherwise some parameters in $p(Y|X)$ are not identifiable. One approach is to impose a parametric model on $p(R|Y, X)$ while $p(Y|X)$ is either parametric or nonparametric; see, e.g., Greenless, Reece, and Zieschang (1982) Robins, Rotnitzky, and Zhao (1994, 1995), Troxel, Lipsitz and Brennam (1997), Troxel, Lipsitz and Harrington (1998), and Qin, Leung, and Shao (2002). We focus on the semiparametric approach in Tang, Little, and Raghunathan (2003) that imposes a parametric model on $p(Y|X)$, assumes that $p(R|Y, X) = p(R|Y)$, but does not require any model on $p(R|Y)$.

As Tang, Little, and Raghunathan (2003) pointed out, their method has a limitation in discarding observed data from sampled subjects with incomplete Y-vectors, which may not be practically desired, and "in some circumstances this information can be incorporated to improve the efficiency of the estimates". One such circumstances is when Y has monotone nonresponse and the nonresponse mechanism is outcome-dependent, i.e.,

$$P(R_t = 1|X, Y, R_1 = \cdots = R_{t-1} = 1) = w_t(Y_t), \quad t = 1, \ldots, T \tag{1.1}$$

(see (17) in Tang, Little, and Raghunathan (2003)), where each w_t is unknown and no model on w_t is required. Under these assumptions, Tang, Little, and Raghunathan (2003) derived estimators based on complete and incomplete data from all sampled subjects.

The purpose of this paper is to derive a method that utilizes all observed data to handle nonmonotone nonresponse under the following counterpart of (1.1) for the nonmonotone nonresponse:

$$P(R_t = 1|X, Y, R_1, \ldots, R_{t-1}) = w_t(Y_t), \quad t = 1, \ldots, T, \tag{1.2}$$

where each w_t is unknown and no model on w_t is required. The method is described in Section 2. It is a semiparametric pseudo-likelihood approach in which $p(Y|X)$ is assumed to be parametric but $w_t(Y_t)$ and $p(X)$ are nonparametric. Properties of the proposed estimators are discussed in Section 3. In Section 4 we apply the proposed method to analyze the HRS data. In Section 5, the finite sample performance of the proposed method is studied by simulation. Some discussion is given in Section 6.

2. Semiparametric Pseudo Likelihoods

For each t, let

$$p(Y_t|X, Y_1, \ldots, Y_{t-1}) = f_t(Y_t|X, Y_1, \ldots, Y_{t-1}, \theta_t), \quad t = 1, \ldots, T, \tag{2.1}$$

where f_t's are known functions and θ_t's are distinct unknown parameter vectors. By factorization,

$$p(Y|X) = \prod_{t=1}^{T} f_t(Y_t|X, Y_1, \ldots, Y_{t-1}, \theta_t).$$

Under monotone nonresponse and assumption (1.1),

$$p(X, Y_1, \ldots, Y_{t-1}|Y_t, R_t = 1) = p(X, Y_1, \ldots, Y_{t-1}|Y_t).$$

Hence, we may use the observed $(X, Y_1, \ldots, Y_t)$ to estimate parameters in $p(X, Y_1, \ldots, Y_{t-1}|Y_t)$. Then, parameters in $p(Y|X)$ can be estimated using

$$p(X, Y_1, \ldots, Y_{t-1}|Y_t) = \frac{p(Y_t|X, Y_1, \ldots, Y_{t-1})p(X, Y_1, \ldots, Y_{t-1})}{\int p(Y_t|x, y_1, \ldots, y_{t-1})p(x, y_1, \ldots, y_{t-1})dxdy_1 \cdots dy_{t-1}}.$$

Under (2.1), Tang, Little, and Raghunathan (2003) proposed to estimate θ_t by maximizing the pseudo-likelihood

$$\prod_{i:R_{i1}=\cdots=R_{it}=1} \frac{f_t(Y_{it}|X_i, Y_{i1}, \ldots, Y_{i(t-1)}, \theta_t)}{\int f_t(Y_{it}|x, y_1, \ldots, y_{t-1}, \theta_t)d\hat{G}(x, y_1, \ldots, y_{t-1})}, \tag{2.2}$$

$t = 1, \ldots, T$, where $(Y_{i1}, \ldots, Y_{iT}, R_{i1}, \ldots, R_{iT}, X_i)$, $i = 1, \ldots, n$, are independent and identically distributed samples from $p(Y, R, X)$, and $\hat{G}(x, y_1, \ldots, y_{t-1})$ is the

empirical distribution based on $(X_i, Y_{i1}, \ldots, Y_{i(t-1)})$ with $R_{i1} = \cdots = R_{i(t-1)} = 1$. In particular, $\hat{G}(x)$ is the empirical distribution based on $X_1, \ldots, X_n$, a nonparametric estimator of the distribution function corresponding to $p(X)$. The likelihood in (2.2) is a pseudo-likelihood because $\hat{G}$ is an estimator. There is a rich literature on the pseudo-likelihood and other related approaches. See, for example, Andersen (1970), Besag (1974), Godambe (1976), and Gong and Samaniego (1981).

Consider nonmonotone nonresponse under assumptions (1.2) and (2.1). Note that maximizing the likelihood given by (2.2) still produces a consistent estimator of θ_t. This can be seen from the fact that, after discarding any observed Y_{it} as long as there is a u such that $u < t$ and Y_{iu} is missing, the nonresponse becomes monotone and (1.1) holds. However, discarding observed data may result in inefficient parameter estimators when the size of the discarded data is appreciable.

We propose a method that utilizes all observed data and estimates θ_t's sequentially. The first step is to obtain an estimate $\hat{\theta}_1$ of θ_1 by maximizing the likelihood in (2.2) with $t = 1$:

$$\prod_{i:R_{i1}=1} g_1(X_i|Y_{i1}, \theta_1), \qquad g_1(X_i|Y_{i1}, \theta_1) = \frac{f_1(Y_{i1}|X_i, \theta_1)}{\int f_1(Y_{i1}|x, \theta_1) d\hat{G}(x)}. \tag{2.3}$$

The second step is to estimate θ_2. The likelihood in (2.2) with $t = 2$ is

$$\prod_{i:R_{i1}=R_{i2}=1} g_2(X_i, Y_{i1}|Y_{i2}, R_{i1} = 1, \theta_2), \tag{2.4}$$

where

$$g_2(X_i, Y_{i1}|Y_{i2}, R_{i1} = 1, \theta_2) = \frac{f_2(Y_{i2}|X_i, Y_{i1}, \theta_2)}{\int f_2(Y_{i2}|x, y_1, \theta_2) d\hat{G}(x, y_1)}$$

is an estimated $p(X_i, Y_{i1}|Y_{i2}, R_{i1} = R_{i2} = 1) = p(X_i, Y_{i1}|Y_{i2}, R_{i1} = 1)$ with θ_2 being fixed. The likelihood in (2.4) does not include any observed Y_{i2} with missing Y_{i1}. To utilize these data, we consider the fact that

$$\begin{aligned} p(X|Y_2, R_2 = 1) &= p(X|Y_2) \\ &= \frac{p(Y_2|X)p(X)}{\int p(Y_2|x)p(x)dx} \\ &= \frac{\left[\int p(Y_2|X, Y_1)p(Y_1|X)dY_1\right] p(X)}{\int \left[\int p(Y_2|x, y_1)p(y_1|x)dy_1\right] p(x)dx} \\ &= \frac{\left[\int f_2(Y_2|X, Y_1, \theta_2)f_1(Y_1|X, \theta_1)dY_1\right] p(X)}{\int \left[\int f_2(Y_2|x, y_1, \theta_2)f_1(y_1|x, \theta_1)dy_1\right] p(x)dx}, \end{aligned}$$

where the first equality follows from (1.2) and the last equality follows from (2.1). With $p(x)$ estimated by $\hat{G}(x)$ and θ_1 estimated by $\hat{\theta}_1$ in the first step, we estimate $p(X_i|Y_{i2}, R_{i2}=1) = p(X_i|Y_{i2})$ (with θ_2 being fixed) by

$$g_2(X_i|Y_{i2},\theta_2) = \frac{\int f_2(Y_{i2}|X_i, y_1, \theta_2) f_1(y_1|X_i, \hat{\theta}_1) dy_1}{\int \left[\int f_2(Y_{i2}|x, y_1, \theta_2) f_1(y_1|x, \hat{\theta}_1) dy_1\right] d\hat{G}(x)}$$

and obtain $\hat{\theta}_2$ by maximizing the pseudo-likelihood

$$\prod_{i:R_{i1}=R_{i2}=1} g_2(X_i, Y_{i1}|Y_{i2}, R_{i1}=1, \theta_2) \prod_{i:R_{i2}=1} g_2(X_i|Y_{i2},\theta_2). \tag{2.5}$$

Note that the second product in (2.5) includes subjects with observed Y_{i2} and missing Y_{i1}. It also includes subjects with observed Y_{i1} and Y_{i2} that have already been included in the first product in (2.5), since the likelihood based on a subject with $R_{i1}=0$ and $R_{i2}=1$, $p(X_i|Y_{i2}, R_{i1}=0, R_{i2}=1)$, may be different from $p(X_i|Y_{i2}, R_{i2}=1)$ under (1.2). Although we can estimate $p(X_i|Y_{i2}, R_{i2}=1)$ by $g_2(X_i|Y_{i2},\theta_2)$ as discussed, we are not able to estimate $p(X_i|Y_{i2}, R_{i1}=0, R_{i2}=1)$ under (1.2) and (2.1).

Having $\hat{\theta}_1, \ldots, \hat{\theta}_{t-1}$, at step $t \le T$ we consider the estimation of θ_t. Let $\mathcal{S}_t$ be the collection of all subsets of $\{1, \ldots, t\}$, including the empty set $\emptyset$. The estimator $\hat{\theta}_t$ of θ_t is then obtained by maximizing the likelihood

$$\prod_{A\in\mathcal{S}_{t-1}} \left[\prod_{i:R_{it}=1, R_{ij}=1, j\in A} g_t(X_i, Y_{ij}, j\in A|Y_{it}, R_{ij}=1, j\in A, \theta_t)\right], \tag{2.6}$$

where, for any subset $A \in \mathcal{S}_{t-1}$, $g_t(X_i, Y_{ij}, j\in A|Y_{it}, R_{ij}=1, j\in A, \theta_t)$ is an estimate of $p(X, Y_j, j\in A|Y_t, R_j=1, j\in A, R_t=1) = p(X, Y_j, j\in A|Y_t, R_j=1, j\in A)$ with θ_t being fixed. The likelihoods in (2.3) and (2.5) are special cases of (2.6) with $t=1$ and 2. The function $g_t(X_i, Y_{ij}, j\in A|Y_{it}, R_{ij}=1, j\in A, \theta_t)$ is obtained as follows. Under assumption (1.2), $p(X, Y_j, j\in A|Y_t, R_j=1, j\in A)$ is equal to

$$\frac{p(Y_t|X, Y_j, j\in A) p(X, Y_j, j\in A|R_j=1, j\in A)}{\int p(Y_t|x, y_j, j\in A) p(x, y_j, j\in A|R_j=1, j\in A) dx \prod_{j\in A} dy_j},$$

which can be replaced by

$$\frac{p(Y_t|X, Y_j, j\in A)}{\int p(Y_t|x, y_j, j\in A) d\hat{G}(x, y_j, j\in A)},$$

where $\hat{G}(X, Y_j, j\in A)$ is the empirical distribution based on $(X_i, Y_{ij}, j\in A)$ with $R_{ij}=1, j\in A$. Using the notation $Y_A = (Y_j, j\in A)$, $y_{A^c} = (y_k, k\notin A)$, and

$dy_{A^c} = \prod_{k \notin A} dy_k$, we obtain that

$$\begin{aligned} p(Y_t|X, Y_j, j \in A) &= \int p(Y_t|X, Y_A, y_{A^c}) p(y_{A^c}|X, Y_A) dy_{A^c} \\ &= \int p(Y_t|X, Y_A, y_{A^c}) \frac{p(Y_A, y_{A^c}|X)}{p(Y_A|X)} dy_{A^c} \\ &= \frac{\int p(Y_t|X, Y_A, y_{A^c}) p(Y_A, y_{A^c}|X) dy_{A^c}}{\int p(Y_A, y_{A^c}|X) dy_{A^c}} \\ &= \frac{\int f_t(Y_t|X, Y_A, y_{A^c}, \theta_t) f_{t-1}(Y_A, y_{A^c}|X, \theta_1, \ldots, \theta_{t-1}) dy_{A^c}}{\int f_{t-1}(Y_A, y_{A^c}|X, \theta_1, \ldots, \theta_{t-1}) dy_{A^c}}, \end{aligned}$$

where, by factorization and (2.1), $f_{t-1}(Y_A, y_{A^c}|X, \theta_1, \ldots, \theta_{t-1})$ can be obtained through

$$f_{t-1}(Y_1, \ldots, Y_{t-1}|X, \theta_1, \ldots, \theta_{t-1}) = \prod_{u=1}^{t-1} f_u(Y_u|X, Y_1, \ldots, Y_{u-1}, \theta_u), \ t = 1, \ldots, T.$$

After replacing θ_u by the estimate $\hat{\theta}_u$ obtained from the previous steps, $u = 1, \ldots, t-1$, we obtain that $g_t(X_i, Y_{ij}, j \in A|Y_{it}, R_{ij} = 1, j \in A, \theta_t)$ is proportional to (after ignoring a factor that does not depend on θ_t)

$$\frac{\int f_t(Y_{it}|X_i, Y_{iA}, y_{A^c}, \theta_t) f_{t-1}(Y_{iA}, y_{A^c}|X_i, \hat{\theta}_1, \ldots, \hat{\theta}_{t-1}) dy_{A^c}}{\int \left[\int f_t(Y_{it}|x, y_A, y_{A^c}, \theta_t) f_{t-1}(y_A, y_{A^c}|x, \hat{\theta}_1, \ldots, \hat{\theta}_{t-1}) dy_{A^c} \right] d\hat{G}(x, y_A)}. \tag{2.7}$$

For some parametric functions f_t in (2.1), (e.g., $p(Y|X)$ is multivariate normal), some integrals in (2.7) can be explicitly worked out. Otherwise, numerical integration is needed.

The likelihood in (2.6) is semiparametric because no model is imposed on w_t in (1.2) and $p(X, Y_j, j \in A|R_j = 1, j \in A)$ is estimated by the nonparametric empirical distribution $\hat{G}(X, Y_j, j \in A)$. It is a pseudo-likelihood since estimators $\hat{G}(X, Y_j, j \in A)$ and $\hat{\theta}_1, \ldots, \hat{\theta}_{t-1}$ are used.

Finally, we explain why we estimate θ_t's sequentially. Note that $\theta_1, \ldots, \theta_{t-1}$ are involved in $p(Y_t|X, Y_j, j \in A)$, which is part of the likelihood using all observed data at time point t. If we do not substitute $\theta_1, \ldots, \theta_{t-1}$ by their estimators from the previous steps, then we have to solve a high-dimensional maximization problem over the parameter $\boldsymbol{\vartheta}_t = (\theta_1, \ldots, \theta_t)$ when t is not small. For example, if $p(Y|X)$ is multivariate normal, then θ_t has dimension $t+2$ and $\boldsymbol{\vartheta}_t$ has dimension $2t+t(t+1)/2$. The sequential procedure does not have this computational burden since maximizing (2.6) is over θ_t with a much smaller dimension than $\boldsymbol{\vartheta}_t$.

3. Properties of Estimators

Asymptotic properties (as $n \to \infty$) of the proposed estimators $\hat{\theta}_t$, $t = 1, \ldots, T$, are studied in this section. The following lemma gives a condition under which the parameter θ_t can be identified.

Lemma 1. *Assume* (1.2) *and* (2.1). *Let* $\theta_1^0, \ldots, \theta_T^0$ *be the true parameter values of* $\theta_1, \ldots, \theta_T$ *in* (2.1). *For any fixed* $t \le T$, θ_t, *and an arbitrary function* $c(Y_t)$, *let*

$$D_{t,\theta_t} = \left\{ Y_t : \frac{f_t(Y_t|X, Y_1, \ldots, Y_{t-1}, \theta_t)}{f_t(Y_t|X, Y_1, \ldots, Y_{t-1}, \theta_t^0)} = c(Y_t) \text{ for any } X, Y_1, \ldots, Y_{t-1} \right\}.$$

If

$$P(R_t = 1|Y_t) > 0 \qquad \text{and} \qquad P(D_{t,\theta_t}) < 1 \quad \text{for any } \theta_t \neq \theta_t^0, \tag{3.1}$$

then, for any $\theta_t \neq \theta_t^0$,

$$E_0[h(\theta_t, \boldsymbol{\vartheta}_{t-1}^0, F^0)] - E_0[h(\theta_t^0, \boldsymbol{\vartheta}_{t-1}^0, F^0)] < 0, \tag{3.2}$$

where $\boldsymbol{\vartheta}_{t-1} = (\theta_1, \ldots, \theta_{t-1})$, E_0 *is the expectation with respect to the true* $p(Y, X)$,

$$h(\theta_t, \boldsymbol{\vartheta}_{t-1}, F) = \sum_{A \in \mathcal{S}_{t-1}} I(R_t = 1, R_j = 1, j \in A) \log p(X, Y_j, j \in A | Y_t, R_j = 1, j \in A),$$

$I(\cdot)$ *is the indicator function, and* F^0 *is the true distribution of* X.

The proof of Lemma 1 and the following theorem are given in the Appendix.

Theorem 1. *Assume the conditions in Lemma* 1 *and that* $E_0(\partial h(\theta_t^0, \boldsymbol{\vartheta}_{t-1}^0, F^0)/\partial\theta_t \partial\theta_t')$ *is positive definite. Assume also that* f_t *in* (2.1) *is continuously twice differentiable with respect to* θ_t *in a neighborhood of* θ_t^0 *and that there exists a function* $M(Y)$ *such that* $E_0[M(Y)] < \infty$ *and*

$$\left\| \frac{\partial}{\partial \boldsymbol{\vartheta}_{t-1}} \int p(Y, X) dX \right\| \le M(Y)$$

for $\boldsymbol{\vartheta}_{t-1}$ *in an open neighborhood of* $\boldsymbol{\vartheta}_{t-1}^0$, $t = 1, \ldots, T$. *Then, as* $n \to \infty$,

$$\sqrt{n}(\hat{\theta}_t - \theta_t^0) \to N(0, \Sigma_t) \quad \text{in distribution} \tag{3.3}$$

for some covariance matrix Σ_t, $t = 1, \ldots, T$.

Although we use all observed data, maximizing (2.6) may not always produce a more efficient estimator than maximizing (2.2), because we may pay a

price in estimating the unknown quantities in (2.6) but not (2.2). It is reasonable to expect that our proposed method is more efficient when the size of the subjects with nonmonotone nonresponse is appreciable. This is supported by our simulation result in Section 5.

For the purpose of inference, a consistent estimator of Σ_t is required. The matrix Σ_t in (3.3) is rather complicated (see the Appendix), because estimators in the previous steps are used in the likelihood for time point t. Thus, a direct substitution estimator is not easy to obtain. We suggest a bootstrap variance estimator $\hat{\Sigma}_t$ obtained by (i) taking B independent bootstrap samples, each of which is a simple random sample with replacement from the n subjects, (ii) computing $\hat{\theta}_t^{*b}$ based on the bth bootstrap sample using the same procedure as that for $\hat{\theta}_t$, $b = 1, \ldots, B$, and (iii) calculating $\hat{\Sigma}_t$ as the sample covariance matrix of $\hat{\theta}_t^{*1}, \ldots, \hat{\theta}_t^{*B}$. Since $\hat{\theta}_t$ is obtained by solving likelihood equations from the likelihood in (2.6), the bootstrap analog $\hat{\theta}_t^{*b}$ is obtained by solving the same likelihood equations with (Y_i, X_i, R_i) replaced by the bootstrap sample (Y_i^*, X_i^*, R_i^*). Thus, $\hat{\Sigma}_t$ is a consistent estimator of the asymptotic covariance matrix of $\hat{\theta}_t$.

4. An Example

The Health and Retirement Study (HRS) of about 22,000 Americans over the age of 50 and their spouses is conducted by the University of Michigan (see more details at the website `http://hrsonline.isr.umich.edu/`). The study is a biannual longitudinal household survey conducted from 1992 to 2006. It paints an emerging portrait of an aging America's physical and mental health, insurance coverage, financial status, family support systems, labor market status, and retirement planning.

The dataset we considered is a cleaned, easy-to-use streamlined public version of the HRS data produced by the Research and Development Corporation (RAND). It is available at `http://hrsonline.isr.umich.edu/meta/rand/index.html`. We used a subset from the original dataset that contains 19,043 households and each household's income at five different years. Missing values exist and nonresponse is nonmonotone. Percentages of data in various nonresponse patterns are shown in Table 1. The percentages of missing data are quite large in this example, partly due to the fact that the household income is the total of several components (e.g., stocks, pensions, and annuities) and the total income is treated as a missing value if any one of these components is missing. Since this is a longitudinal survey, the percentage of households with no missing value (the second last column of Table 1) is small and decreases as t increases.

To illustrate how to apply the procedure in Section 2 we assume a parametric model on $p(Z_1, \ldots, Z_5|X)$, where Z_t is the total household income in year t and X is the number of years of education treated as a covariate that ranges from 0 to 17

Table 1. Percentage of HRS Household Income Data in Different Patterns.

	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	% of data	% of data used in likelihood (2.2)	% of data used in likelihood (2.6)
$t=1$	1					32.85	32.85	32.85
	0					67.15		
	0	1				10.96		
$t=2$	1	1				18.03	18.03	28.99
	x	0				71.01		
	0	0	1			4.96		
	0	1	1			4.85		
$t=3$	1	0	1			4.46	10.42	24.69
	1	1	1			10.42		
	x	x	0			75.31		
	0	0	0	1		9.90		
	0	0	1	1		2.24		
	0	1	0	1		1.81		
	0	1	1	1		2.82		
$t=4$	1	0	0	1		2.03	6.99	31.12
	1	0	1	1		2.56		
	1	1	0	1		2.78		
	1	1	1	1		6.99		
	x	x	x	0		68.88		
	0	0	0	0	1	4.17		
	0	0	0	1	1	5.28		
	0	0	1	0	1	0.74		
	0	0	1	1	1	1.14		
	0	1	0	0	1	0.70		
	0	1	0	1	1	0.82		
	0	1	1	0	1	0.65		
	0	1	1	1	1	1.69		
$t=5$	1	0	0	0	1	0.98	4.87	27.94
	1	0	0	1	1	1.00		
	1	0	1	0	1	0.63		
	1	0	1	1	1	1.50		
	1	1	0	0	1	0.89		
	1	1	0	1	1	1.63		
	1	1	1	0	1	1.26		
	1	1	1	1	1	4.87		
	x	x	x	x	0	72.06		

0: the outcome is missing.
1: the outcome is observed.
x: 0 or 1.

with mean 12.74. Since Z_t has a skewed distribution in this example, we applied the inverse hyperbolic sine transformation that was used by the RAND to impute

the missing household income, $Y_t = \log(Z_t + \sqrt{1+Z_t^2})$. The transformation is close to log transformation when Z_t is large. We assume that $p(Y_1, \ldots, Y_5|X)$ is multivariate normal:

$$Y_t = \beta_{t0} + \beta_{t1}X + \beta_{t2}Y_1 + \cdots + \beta_{t(t-1)}Y_{t-1} + \varepsilon_t, \quad t = 1, \ldots, 5, \tag{4.1}$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma_t^2)$ and ε_t's are independent.

For comparison, we applied three approaches for parameter estimation: (i) the method of using data from subjects without any missing value, i.e., ignoring subjects with incomplete data, (ii) the approach of maximizing likelihood (2.2), and (iii) the proposed approach of maximizing likelihood (2.6). Method (i) is justified when nonresponse is covariate-dependent, i.e., $p(R|Y, X) = p(R|X)$. For each approach, we applied the bootstrap method with $B = 200$ for estimating the standard deviations of the estimates. In this example, it is of interest to estimate the mean household income at each year, in addition to the β- and σ-parameters. We obtained estimates of the mean household income $E(Z_t)$ based on the estimated parameters in $p(Y_1, \ldots, Y_5|X)$ and the inverse of the transformation $Y_t = \log(Z_t + \sqrt{1+Z_t^2})$. All parameter estimates and their estimated standard deviations are given in Table 2.

It can be seen from Table 2 that the estimates obtained by ignoring subjects with incomplete data are very different from those obtained by handling missing data. In terms of the mean household income, ignoring subjects with incomplete data results in negatively biased estimates, which indicates that household income of a nonrespondent is typically higher than the household income of a respondent. An interesting observation is that the household income estimate based on methods of handling missing data has two significant decreases, one at 1999 and the other at 2003, probably because this is the household income of retired seniors that depends on the fluctuation of the stock market. On the other hand, the household income estimate obtained by ignoring subjects with incomplete data does not follow this pattern.

Comparing the two methods of handling nonrespondents, we find that they provide comparable mean estimates (they are the same at $t=1$, year 1997), but the proposed method of maximizing (2.6) has much smaller bootstrap estimated standard deviation than the method of maximizing (2.2) when $t > 1$. Although we do not know the truth, the simulation results in Section 5 suggest that our proposed method is better. In this example, discarding nonmonotone missing values has a large effect: the percentage of data used in likelihood (2.2) decreases to 11% when $t \geq 3$ and becomes as small as 4.87% when $t = 5$, whereas the percentage of data used in likelihood (2.6) is between 24.69% and 32.85% for all t.

Table 2. Empirical Results For HRS Example Parameter estimate (Standard error based on 200 bootstrap samples).

parameter	ignoring incomplete	method maximizing likelihood (2.2)	maximizing likelihood (2.6)
β_{10}	8.845 (0.202)	9.568 (0.154)	9.568 (0.154)
β_{11}	0.176 (0.016)	0.136 (0.015)	0.136 (0.015)
σ_1	1.579 (0.146)	1.370 (0.116)	1.370 (0.116)
β_{20}	5.161 (0.709)	6.199 (0.826)	6.724 (0.638)
β_{21}	0.104 (0.015)	0.199 (0.023)	0.177 (0.017)
β_{22}	0.394 (0.074)	0.219 (0.058)	0.201 (0.047)
σ_2	1.530 (0.140)	1.780 (0.211)	1.748 (0.171)
β_{30}	3.088 (0.775)	3.140 (1.952)	5.428 (1.114)
β_{31}	0.064 (0.015)	0.214 (0.029)	0.156 (0.020)
β_{32}	0.264 (0.078)	0.222 (0.091)	0.143 (0.045)
β_{33}	0.371 (0.090)	0.279 (0.081)	0.201 (0.049)
σ_3	1.382 (0.133)	2.047 (0.341)	1.802 (0.233)
β_{40}	2.662 (0.617)	3.194 (1.859)	3.157 (1.182)
β_{41}	0.045 (0.011)	0.106 (0.027)	0.105 (0.022)
β_{42}	0.186 (0.062)	0.155 (0.078)	0.186 (0.061)
β_{43}	0.173 (0.064)	0.155 (0.063)	0.132 (0.038)
β_{44}	0.350 (0.091)	0.285 (0.081)	0.271 (0.067)
σ_4	0.909 (0.089)	1.187 (0.252)	1.571 (0.226)
β_{50}	3.133 (0.815)	4.057 (2.556)	4.280 (1.101)
β_{51}	0.052 (0.013)	0.108 (0.031)	0.108 (0.025)
β_{52}	0.098 (0.054)	0.067 (0.065)	0.065 (0.046)
β_{53}	0.040 (0.039)	0.051 (0.043)	0.071 (0.034)
β_{54}	0.130 (0.056)	0.122 (0.055)	0.138 (0.045)
β_{55}	0.385 (0.100)	0.268 (0.187)	0.214 (0.072)
σ_5	0.922 (0.110)	1.041 (0.263)	1.259 (0.193)
mean household income			
1997 ($t = 1$)	19218 (1012)	35338 (2361)	35338 (2361)
1999 ($t = 2$)	19021 (915)	29516 (2278)	31702 (2053)
2001 ($t = 3$)	19115 (926)	36956 (4626)	32548 (2825)
2003 ($t = 4$)	21556 (893)	32134 (3515)	28047 (1675)
2005 ($t = 5$)	21608 (739)	28737 (2244)	27871 (1155)

5. Simulation Results

A simulation study was conducted with $T = 5$, $n = 700$, and model (4.1). The covariate X was generated from the Poisson distribution with mean 12, but this information was not used in the estimation procedure. The true parameter values in the simulation were similar to estimated values in the HRS example presented in Section 4:

$$\theta_1=(\beta_{10}, \beta_{11}, \sigma_1) = (9.57, 0.14, 1.34),$$
$$\theta_2=(\beta_{20}, \beta_{21}, \beta_{22}, \sigma_2) = (6.72, 0.18, 0.20, 1.73),$$
$$\theta_3=(\beta_{30}, \beta_{31}, \beta_{32}, \beta_{33}, \sigma_3) = (5.43, 0.16, 0.14, 0.20, 1.73),$$
$$\theta_4=(\beta_{40}, \beta_{41}, \beta_{42}, \beta_{43}, \beta_{44}, \sigma_4) = (3.16, 0.10, 0.19, 0.13, 0.27, 1.41),$$
$$\theta_5=(\beta_{50}, \beta_{51}, \beta_{52}, \beta_{53}, \beta_{54}, \beta_{55}, \sigma_5) = (4.28, 0.11, 0.06, 0.07, 0.14, 0.21, 1.26).$$

The nonresponse indicators were generated from a logistic model with

$$P(R_t = 1|X, Y, R_j, j \neq t) = \frac{\exp(40 - 3.59Y_t)}{1 + \exp(40 - 3.59Y_t)}$$

for $t = 1, \ldots, 5$. The expected rate of missing data for each t is about 50%. Since components of Y are missing independently, the ratio of the number of subjects included in likelihood (2.2) over the number of subjects included in likelihood (2.6) is about $2^{-(t-1)}$.

For comparison, we considered the four methods of estimating θ_t's (i) using all data as a standard; (ii) using data from subjects without nonresponse and ignoring subjects with incomplete data; (iii) maximizing likelihood (2.2) by discarding observed data after first nonresponse; and (iv) maximizing likelihood (2.6).

Based on 1000 simulation runs, empirical results for the four quantities are given in Table 3: the bias and standard deviation (SD) of estimators of all parameters, and the coverage probability (CP) and length (CL) of the approximate 95% confidence interval = estimator $\pm$ $1.96\sqrt{\text{the bootstrap variance estimate}}$, where $B = 100$. The results in Table 3 can be summarized as follows.

1. The estimator obtained by ignoring subjects with incomplete data can be seriously biased, although in a few cases the bias is not large. The coverage probability of the related confidence interval is much lower than the nominal level 95% when the absolute bias is much larger than the standard deviation.
2. Maximizing (2.2) and (2.6) produce almost unbiased estimators and coverage probabilities close to 95%, except for the case of σ_5. In a few cases, maximizing (2.2) is somewhat too conservative.

Table 3. Empirical Results Based on 1000 Runs.

		method			
parameter	quantity	standard (no missing)	ignoring incomplete	maximizing likelihood (2.2)	maximizing likelihood (2.6)
β_{10}	bias (SD)	-0.005 (0.18)	-0.252 (0.40)	-0.103 (0.84)	-0.103 (0.84)
	CP (CL)	94.8 (0.71)	88.6 (1.52)	95.7 (4.06)	95.7 (4.06)
β_{11}	bias (SD)	0.000 (0.01)	-0.083 (0.04)	0.014 (0.08)	0.014 (0.08)
	CP (CL)	95.7 (0.06)	43.9 (0.15)	95.8 (0.44)	95.8 (0.44)
σ_1	bias (SD)	-0.001 (0.04)	-0.382 (0.08)	0.049 (0.47)	0.049 (0.47)
	CP (CL)	93.9 (0.14)	1.3 (0.3)	95.8 (3.09)	95.8 (3.09)
β_{20}	bias (SD)	-0.019 (0.53)	0.856 (1.56)	-0.516 (2.70)	-0.169 (1.83)
	CP (CL)	94.3 (2.04)	89.5 (5.98)	97.0 (15.16)	94.7 (8.59)
β_{21}	bias (SD)	0.001 (0.02)	-0.099 (0.05)	0.028 (0.12)	0.014 (0.08)
	CP (CL)	94.8 (0.08)	52.8 (0.21)	97.4 (0.76)	97.0 (0.40)
β_{22}	bias (SD)	0.001 (0.05)	-0.104 (0.15)	0.032 (0.21)	0.008 (0.15)
	CP (CL)	93.6 (0.19)	89.1 (0.60)	98.0 (1.32)	93.4 (0.71)
σ_2	bias (SD)	-0.001 (0.05)	-0.458 (0.10)	0.120 (0.77)	0.064 (0.47)
	CP (CL)	94.2 (0.18)	2.7 (0.40)	95.6 (5.38)	95.7 (2.74)
β_{30}	bias (SD)	-0.031 (0.58)	1.445 (1.82)	-1.288 (4.69)	-0.191 (2.02)
	CP (CL)	94.3 (2.27)	84.7 (7.05)	98.2 (22.1)	95.3 (9.23)
β_{31}	bias (SD)	0.001 (0.02)	-0.082 (0.05)	0.047 (0.19)	0.011 (0.06)
	CP (CL)	95.1 (0.08)	67.4 (0.22)	97.9 (0.86)	96.4 (0.33)
β_{32}	bias (SD)	0.001 (0.05)	-0.073 (0.16)	0.051 (0.34)	0.009 (0.15)
	CP (CL)	94.6 (0.19)	91.4 (0.61)	98.8 (1.67)	95.1 (0.66)
β_{33}	bias (SD)	0.001 (0.04)	-0.102 (0.12)	0.041 (0.26)	0.001 (0.12)
	CP (CL)	93.3 (0.15)	85.1 (0.46)	97.9 (1.43)	94.4 (0.56)
σ_3	bias (SD)	-0.001 (0.05)	-0.434 (0.11)	0.225 (1.19)	0.040 (0.41)
	CP (CL)	93.4 (0.18)	5.3 (0.41)	95.7 (5.75)	95.2 (2.29)

Bias: bias of estimator
SD: standard deviation of estimator
CP: coverage probability of confidence interval in %
CL: length of confidence interval

3. In terms of the efficiency (the standard deviation of the estimator and the length of confidence interval), the proposed method of maximizing (2.6) is better than the method of maximizing (2.2). The improvement is more substantial when t is large.

6. Discussion

In the presence of nonmonotone missing outcomes in longitudinal data, we propose an estimation procedure that utilizes all observed data. At time point t, the likelihood containing all observed data under nonmonotone nonresponse involves parameters in the likelihoods for previous time points. Thus, we propose a computationally sensible sequential procedure that substitutes these parameters by their estimators obtained at time points $1, \ldots, t-1$ and then maximizes the

Table 3. (continued)

		method			
parameter	quantity	standard (no missing)	ignoring incomplete	maximizing likelihood (2.2)	maximizing likelihood (2.6)
β_{40}	bias (SD)	0.012 (0.53)	1.628 (1.77)	-0.195 (3.12)	0.052 (1.62)
	CP (CL)	94.3 (1.97)	82.7 (7.03)	96.1 (14.69)	94.7 (6.54)
β_{41}	bias (SD)	-0.001 (0.02)	-0.038 (0.05)	0.003 (0.07)	-0.002 (0.04)
	CP (CL)	93.6 (0.07)	88.1 (0.20)	96.8 (0.36)	96.3 (0.16)
β_{42}	bias (SD)	-0.001 (0.04)	-0.060 (0.14)	0.004 (0.19)	-0.002 (0.12)
	CP (CL)	94.1 (0.16)	92.4 (0.55)	97.1 (0.97)	94.3 (0.47)
β_{43}	bias (SD)	-0.000 (0.03)	-0.040 (0.11)	0.004 (0.15)	0.000 (0.10)
	CP (CL)	94.4 (0.12)	90.9 (0.42)	96.4 (0.72)	93.9 (0.36)
β_{44}	bias (SD)	0.000 (0.03)	-0.086 (0.10)	0.006 (0.17)	-0.003 (0.10)
	CP (CL)	94.8 (0.12)	84.9 (0.42)	96.5 (0.83)	94.4 (0.40)
σ_4	bias (SD)	-0.002 (0.04)	-0.253 (0.09)	-0.034 (0.46)	-0.019 (0.20)
	CP (CL)	93.6 (0.15)	24.1 (0.35)	93.3 (2.20)	93.0 (0.88)
β_{50}	bias (SD)	-0.017 (0.47)	1.247 (1.73)	-0.076 (3.16)	0.048 (1.28)
	CP (CL)	94.4 (1.82)	88.2 (6.67)	95.0 (12.55)	96.4 (5.35)
β_{51}	bias (SD)	-0.000 (0.02)	-0.036 (0.05)	0.002 (0.08)	-0.001 (0.03)
	CP (CL)	94.1 (0.06)	87.2 (0.18)	95.5 (0.36)	95.8 (0.14)
β_{52}	bias (SD)	-0.001 (0.04)	-0.019 (0.13)	-0.003 (0.19)	-0.001 (0.10)
	CP (CL)	95.1 (0.14)	94.3 (0.50)	97.2 (0.82)	94.8 (0.41)
β_{53}	bias (SD)	0.001 (0.03)	-0.024 (0.10)	-0.003 (0.15)	-0.001 (0.08)
	CP (CL)	94.2 (0.11)	94.2 (0.38)	96.5 (0.64)	93.3 (0.32)
β_{54}	bias (SD)	0.002 (0.03)	-0.037 (0.10)	0.004 (0.18)	-0.000 (0.08)
	CP (CL)	95.6 (0.11)	91.7 (0.38)	96.0 (0.67)	93.9 (0.33)
β_{55}	bias (SD)	-0.001 (0.03)	-0.056 (0.10)	0.003 (0.17)	-0.002 (0.09)
	CP (CL)	94.6 (0.13)	90.5 (0.42)	97.1 (0.76)	96.0 (0.39)
σ_5	bias (SD)	0.000 (0.03)	-0.221 (0.08)	-0.050 (0.49)	-0.029 (0.16)
	CP (CL)	93.4 (0.13)	26.8 (0.32)	89.1 (1.70)	89.3 (0.66)

Bias: bias of estimator
SD: standard deviation of estimator
CP: coverage probability of confidence interval in %
CL: length of confidence interval

resulting pseudo-likelihood at time point t. Our method is semiparametric, i.e., a parametric model on the likelihood of the outcome given covariates is imposed but no model on the nonresponse mechanism is required.

To relax Assumption (1.2), suppose that, in addition to the covariate X, there is a fully observed discrete covariate Z, which can be multivariate or longitudinal. Assume

$$P(R_t = 1|X, Z, Y, R_1, \ldots, R_{t-1}) = w_t(Y_t, Z), \quad t = 1, \ldots, T, \tag{6.1}$$

where w_t is still nonparametric. The counterpart of Assumption (2.1) is

$$p(Y_t|X, Z, Y_1, \ldots, Y_{t-1}) = f_t(Y_t|X, Z, Y_1, \ldots, Y_{t-1}, \theta_t), \quad t = 1, \ldots, T. \tag{6.2}$$

Note that

$$p(X, Y_A|Y_t, Z, R_j = 1, j \in A) = \frac{p(Y_t|X, Z, Y_A)p(X, Y_A|Z, R_j = 1, j \in A)}{\int p(Y_t|x, y_A, Z)p(x, y_A|Z, R_j = 1, j \in A)dxdy_A},$$

where $Y_A = (Y_j, j \in A)$ and y_A is a realization of Y_A. The method described in Section 2 can be modified to maximizing

$$\prod_z \prod_{A \in \mathcal{S}_{t-1}} \left[\prod_{i:R_{it}=1, R_{ij}=1, j \in A} g_t(X_i, Y_{ij}, j \in A|Y_{it}, Z_i = z, R_{ij} = 1, j \in A, \theta_t) \right],$$

where $g_t(X_i, Y_{ij}, j \in A|Y_{it}, Z_i = z, R_{ij} = 1, j \in A, \theta_t)$ is proportional to

$$\frac{\int f_t(Y_{it}|X_i, Z_i = z, Y_{iA}, y_{A^c}, \theta_t) f_{t-1}(Y_{iA}, y_{A^c}|X_i, Z_i = z, \hat{\theta}_1, \ldots, \hat{\theta}_{t-1}) dy_{A^c}}{\int [\int f_t(Y_{it}|x, Z_i = z, y_A, y_{A^c}, \theta_t) f_{t-1}(y_A, y_{A^c}|x, Z_i = z, \hat{\theta}_1, \ldots, \hat{\theta}_{t-1}) dy_{A^c}] d\hat{G}_z(x, y_A)}$$

and $\hat{G}_z(x, y_A)$ is the empirical distribution based on $(X_i, Y_{ij}, j \in A)$ with $R_{ij} = 1$, $j \in A$, and $Z_i = z$.

One of the two key assumptions required for our approach is the outcome-dependent nonresponse Assumption (1.2) or (6.1), which is a counterpart of Assumption (1.1) in Tang, Little, and Raghunathan (2003), although no model is required for the function w_t. Unfortunately, no assumption on the nonresponse mechanism, such as (1.1), (1.2), (6.1), the ignorable nonresponse assumption, or the covariate-dependent nonresponse assumption, can be checked using data due to the presence of missing values. We may be able to show that the ignorable nonresponse assumption does not hold under some nonignorable nonresponse assumption (such as (1.2)) which itself cannot be checked using data. However, this does not lower the importance of studying valid estimation methods under (1.2) or (6.1) as well as other assumptions. Our investigation will be continued. The results will be useful if the assumption is appropriate or if one would like to perform a sensitivity analysis under different assumptions.

Note that we need to assume at least one of $p(Y|X)$ and $p(R|Y, X)$ is parametric to be able to identify parameters in $p(Y|X)$. The second key assumption for our approach is the parametric model (2.1) or (6.2). Parametric models are sensitive to model violations, and we have to be careful when we apply them. The same issue exists for the likelihood approach in Little and Rubin (2002) under ignorable nonresponse. With nonignorable nonresponse or the general ignorable nonresponse, unfortunately, we are not able to verify parametric models using observed data. The parametric model is imposed on the density $p(Y_t|X, Y_1, \ldots, Y_{t-1})$, which is a mixture of $p(Y_t|X, Y_1, \ldots, Y_{t-1}, R_t = 1)$ and $p(Y_t|X, Y_1, \ldots, Y_{t-1}, R_t = 0)$, and we are not able to check a parametric model assumption on $p(Y_t|X, Y_1, \ldots, Y_{t-1}, R_t = 0)$ since no Y_t-observation comes from

it, although we can check a model on $p(Y_t|X, Y_1, \ldots, Y_{t-1}, R_t = 1)$ using observed data. Under the stronger assumption of covariate-dependent nonresponse, we can check Assumption (2.1) or (6.2) using observed data and the fact that $p(Y|X, R) = p(Y|X)$.

A popular parametric model on $p(Y|X)$ is the multivariate normal model. Other choices of f_t in (2.1) include the generalized linear models by treating X, $Y_1, \ldots, Y_{t-1}$ as covariates at time t. When $X = (X_1, \ldots, X_T)$ is time-dependent, it is sensible to assume that Y_t is statistically related to $X_1, \ldots, X_t$ only, $t = 1, \ldots, T$. In such cases, the proposed method can be modified with X replaced by $X_1, \ldots, X_t$ at each time point t, because (2.1) becomes

$$p(Y_t|X, Y_1, \ldots, Y_{t-1}) = f_t(Y_t|X_1, \ldots, X_t, Y_1, \ldots, Y_{t-1}, \theta_t).$$

Acknowledgements

The authors would like to thank a referee and an associate editor for their helpful comments and suggestions. The research was partially supported by the NSF Grants SES-0705033 and DMS-1007454.

Appendix: Proofs

Proof of Lemma 1. Let G_A^0 be the distribution of $(X, Y_j, j \in A)$ with $R_j = 1$, $j \in A$, and let $\psi_A(\theta_t, \boldsymbol{\vartheta}_{t-1}, G_A) = p(X, Y_j, j \in A|Y_t, R_j = 1, j \in A)$. Then the left hand side of (3.2) is equal to

$$\sum_{A \in \mathcal{S}_{t-1}} E_0 \left[I(R_t = 1, R_j = 1, j \in A) \log \frac{\psi_A(\theta_t, \boldsymbol{\vartheta}_{t-1}^0, G_A^0)}{\psi_A(\theta_t^0, \boldsymbol{\vartheta}_{t-1}^0, G_A^0)} \middle| Y_t, R_j = 1, j \in A \right].$$

Each term in this sum is less or equal to 0 by Jensen's inequality. Since the log-function is strictly concave, this sum is 0 if and only if all terms are 0, which implies that

$$E_0 \left[I(R_1 = \cdots = R_t = 1) \log \frac{\psi_{A_{t-1}}(\theta_t, \boldsymbol{\vartheta}_{t-1}^0, F^0)}{\psi_{A_{t-1}}(\theta_t^0, \boldsymbol{\vartheta}_{t-1}^0, F^0)} \middle| Y_t, R_1 = \cdots = R_{t-1} = 1 \right] = 0,$$

where $A_{t-1} = \{1, \ldots, t-1\}$. By Jensen's inequality again, this implies that either $P(D_{t,\theta_t}) = 1$ or $P(R_t = 1|Y_t, R_1 = \cdots = R_{t-1} = 1) = P(R_t = 1|Y_t) = 0$, which contradicts condition (3.1). This proves (3.2).

Proof of Theorem 1. For any function $f(\theta_t, z)$, where z denotes some other parameters and/or random variables, let $f_{\theta_t}(\theta_t, z) = \frac{\partial}{\partial \theta_t} f(\theta_t, z)$ and $f_{\theta_t \theta_t'}(\theta_t, z) = \frac{\partial}{\partial \theta_t} f_{\theta_t}(\theta_t, z)$. Let G_A denote a distribution on the same space as that of G_A^0, $G = (G_A, A \in \mathcal{S}_{t-1}, t = 1, \ldots, T)$, $G^0 = (G_A^0, A \in \mathcal{S}_{t-1}, t = 1, \ldots, T)$, $\hat{G}_A$ be

the empirical distribution estimator of G_A^0, $\hat{G} = (\hat{G}_A, A \in \mathcal{S}_{t-1}, t = 1, \ldots, T)$, $\psi_A^{(i)}(\theta_t, \boldsymbol{\vartheta}_{t-1}, G_A)$ be $\psi_A(\theta_t, \boldsymbol{\vartheta}_{t-1}, G_A)$ based on data from subject i, $I_{it}(A) = I(R_{it} = 1, R_{ij} = 1, j \in A)$, and

$$l^{(t)}(\theta_t, \boldsymbol{\vartheta}_{t-1}, G) = \frac{1}{n} \sum_{i=1}^{n} \sum_{A \in S_{t-1}} I_{it}(A) \log \psi_A^{(i)}(\theta_t, \boldsymbol{\vartheta}_{t-1}, G_A).$$

We show the consistency of $\hat{\theta}_t$, assuming that we have shown the consistency of $\hat{\theta}_1, \ldots, \hat{\theta}_{t-1}$. Since $(\hat{\boldsymbol{\vartheta}}_{t-1}, \hat{G}_A) \to (\boldsymbol{\vartheta}^0, G_A^0)$ as $n \to \infty$, by the differentiability of $l^{(t)}$ and the Law of Large Numbers, almost surely we have

$$\begin{aligned} l^{(t)}(\theta_t, \hat{\boldsymbol{\vartheta}}_{t-1}, \hat{G}) - l^{(t)}(\theta_t^0, \hat{\boldsymbol{\vartheta}}_{t-1}, \hat{G}) &= l^{(t)}(\theta_t, \boldsymbol{\vartheta}_{t-1}^0, G^0) - l^{(t)}(\theta_t^0, \boldsymbol{\vartheta}_{t-1}^0, G^0) + o(1) \\ &= E_0 \left(h(\theta_t, \boldsymbol{\vartheta}_{t-1}^0, F^0) - h(\theta_t^0, \boldsymbol{\vartheta}_{t-1}^0, F^0) \right) + o(1), \end{aligned}$$

which is negative for sufficiently large n (Lemma 1). If $\lim_{n\to\infty} \hat{\theta}_t \neq \theta_t^0$, then there is a subsequence of $\{\hat{\theta}_t\}$, say $\{\hat{\theta}_{k,t}\}$, such that $\hat{\theta}_{k,t} \to \theta_t^{'}$ and $\theta_t^{'} \neq \theta_t^0$. This implies that $l^{(t)}(\hat{\theta}_{k,t}, \hat{\boldsymbol{\vartheta}}_{t-1}, \hat{G}) < l^{(t)}(\theta_t^0, \hat{\boldsymbol{\vartheta}}_{t-1}, \hat{G})$ when n is large enough. This contradicts to the fact that $\hat{\theta}_t$ is a maximum of $l^{(t)}(\theta, \hat{\boldsymbol{\vartheta}}_{t-1}, \hat{G})$. Hence, $\hat{\theta}_t$ is consistent.

We next show the asymptotic normality of $\hat{\theta}_t$, assuming that $\hat{\boldsymbol{\vartheta}}_{t-1}$ is asymptotically normal. By Taylor's expansion and the fact that $l_{\theta_t}^{(t)}(\hat{\theta}_t, \hat{\boldsymbol{\vartheta}}_{t-1}, \hat{G}) = 0$,

$$\begin{aligned} -l_{\theta_t}^{(t)}(\theta_t^0, \boldsymbol{\vartheta}_{t-1}^0, G^0) &= l_{\theta_t}^{(t)}(\hat{\theta}_t, \hat{\boldsymbol{\vartheta}}_{t-1}, \hat{G}) - l_{\theta_t}^{(t)}(\theta_t^0, \hat{\boldsymbol{\vartheta}}_{t-1}, \hat{G}) \\ &\quad + l_{\theta_t}^{(t)}(\theta_t^0, \hat{\boldsymbol{\vartheta}}_{t-1}, \hat{G}) - l_{\theta_t}^{(t)}(\theta_t^0, \boldsymbol{\vartheta}_{t-1}^0, \hat{G}) \\ &\quad + l_{\theta_t}^{(t)}(\theta_t^0, \boldsymbol{\vartheta}_{t-1}^0, \hat{G}) - l_{\theta_t}^{(t)}(\theta_t^0, \boldsymbol{\vartheta}_{t-1}^0, G^0) \\ &= l_{\theta_t \theta_t'}^{(t)}(\theta_t^0, \hat{\boldsymbol{\vartheta}}_{t-1}, \hat{G})(\hat{\theta}_t - \theta_t^0) + o_p(n^{-1/2}) \\ &\quad + l_{\theta_t \boldsymbol{\vartheta}_{t-1}'}^{(t)}(\theta_t^0, \boldsymbol{\vartheta}_{t-1}^0, \hat{G})(\hat{\boldsymbol{\vartheta}}_{t-1} - \boldsymbol{\vartheta}_{t-1}^0) + o_p(n^{-1/2}) \\ &\quad + B_{tn}, \end{aligned}$$

where

$$\begin{aligned} B_{tn} = \frac{1}{n} \sum_{i=1}^{n} \sum_{A \in \mathcal{S}_{t-1}} I_{it}(A) \Bigg[& \frac{\int \phi_{\theta_t}(Y_{it}, x, y_A, \theta_t^0, \boldsymbol{\vartheta}_{t-1}^0) dG_A^0}{\int \phi(Y_{it}, x, y_A, \theta_t^0, \boldsymbol{\vartheta}_{t-1}^0) dG_A^0} \\ & - \frac{\int \phi_{\theta_t}(Y_{it}, x, y_A, \theta_t^0, \boldsymbol{\vartheta}_{t-1}^0) d\hat{G}_A}{\int \phi(Y_{it}, x, y_A, \theta_t^0, \boldsymbol{\vartheta}_{t-1}^0) d\hat{G}_A} \Bigg], \end{aligned}$$

$$\phi(Y_{it}, x, y_A, \theta_t, \boldsymbol{\vartheta}_{t-1}) = \int f_t(Y_{it} | x, y_A, y_{A^c}, \theta_t) f_{t-1}(y_A, y_{A^c} | x, \boldsymbol{\vartheta}_{t-1}) dy_{A^c}.$$

A direct calculation shows that

$$B_{tn} = \frac{1}{n}\sum_{i=1}^{n}\sum_{A\in\mathcal{S}_{t-1}} I_{it}(A)\left[\frac{\int \phi_{\theta_t}(Y_{it},x,y_A,\theta_t^0,\boldsymbol{\vartheta}_{t-1}^0)d(G_A^0-\hat{G}_A)}{\int \phi(Y_{it},x,y_A,\theta_t^0,\boldsymbol{\vartheta}_{t-1}^0)dG_A^0}\right.$$
$$\left.+\frac{\int \phi_{\theta_t}(Y_{it},x,y_A,\theta_t^0,\boldsymbol{\vartheta}_{t-1}^0)d\hat{G}_A}{\int \phi(Y_{it},x,y_A,\theta_t^0,\boldsymbol{\vartheta}_{t-1}^0)dG_A^0}\frac{\int \phi(Y_{it},x,y_A,\theta_t^0,\boldsymbol{\vartheta}_{t-1}^0)d(\hat{G}_A-G_A^0)}{\int \phi(Y_{it},x,y_A,\theta_t^0,\boldsymbol{\vartheta}_{t-1}^0)d\hat{G}_A}\right]$$
$$= \tilde{B}_{tn} + o_p(n^{-1/2}),$$

where

$$\tilde{B}_{tn} = \frac{1}{n}\sum_{i=1}^{n}\sum_{A\in\mathcal{S}_{t-1}} I_{it}(A)\left[\frac{\int \phi_{\theta_t}(Y_{it},x,y_A,\theta_t^0,\boldsymbol{\vartheta}_{t-1}^0)d(G_A^0-\hat{G}_A)}{\int \phi(Y_{it},x,y_A,\theta_t^0,\boldsymbol{\vartheta}_{t-1}^0)dG_A^0}\right.$$
$$\left.+\frac{\int \phi_{\theta_t}(Y_{it},x,y_A,\theta_t^0,\boldsymbol{\vartheta}_{t-1}^0)dG_A^0\int \phi(Y_{it},x,y_A,\theta_t^0,\boldsymbol{\vartheta}_{t-1}^0)d(\hat{G}_A-G_A^0)}{[\int \phi(Y_{it},x,y_A,\theta_t^0,\boldsymbol{\vartheta}_{t-1}^0)dG_A^0]^2}\right].$$

By the Law of Large Numbers, $l^{(t)}_{\theta_t\theta_t'}(\theta_t^0,\hat{\boldsymbol{\vartheta}}_{t-1},\hat{G})$ converges to $E_0\left(\frac{\partial^2 h(\theta_t^0,\boldsymbol{\vartheta}_{t-1}^0,G^0)}{\partial\theta_t\partial\theta_t'}\right)$ in probability, which is assumed to be positive definite. Combining these results, we obtain that

$$\hat{\theta}_t-\theta_t^0 = -\left[E_0\left(\frac{\partial^2 h(\theta_t^0,\boldsymbol{\vartheta}_{t-1}^0,G^0)}{\partial\theta_t\partial\theta_t'}\right)\right]^{-1}\left[l^{(t)}_{\theta_t}(\theta_t^0,\boldsymbol{\vartheta}_{t-1}^0,G^0)\right.$$
$$\left.+\, l^{(t)}_{\theta_t\boldsymbol{\vartheta}_{t-1}'}(\theta_t^0,\boldsymbol{\vartheta}_{t-1}^0,G^0)(\hat{\boldsymbol{\vartheta}}_{t-1}-\boldsymbol{\vartheta}_{t-1}^0)+\tilde{B}_{tn}\right]+o_p(n^{-1/2}).$$

By induction, we have

$$\sqrt{n}(\hat{\theta}_t-\theta_t^0) = -\left[E_0\left(\frac{\partial h(\theta_t^0,\boldsymbol{\vartheta}_{t-1}^0,F^0)}{\partial\theta_t\partial\theta_t'}\right)\right]^{-1}\sqrt{n}\sum_{s=1}^{t}[C_s^t l^{(s)}_{\theta_s}(\theta_s^0,\boldsymbol{\vartheta}_{s-1}^0,G^0)+\tilde{B}_{sn}]$$
$$+o_p(1),$$

where C_k^t, $k=1,\ldots,t$, are some non-random matrices with $C_t^t = I$. Note that $\tilde{B}_{tn}$ is a type of V-statistic and each $l^{(t)}_{\theta_t}(\theta_t^0,\boldsymbol{\vartheta}_{t-1}^0,G^0)$ is an average of independent random vectors with mean 0. By the Central Limit Theorem, (3.3) holds. The form of Σ_t in (3.3), however, is very complicated because $\tilde{B}_{tn}$ and $l^{(t)}_{\theta_t}(\theta_t^0,\boldsymbol{\vartheta}_{t-1}^0,G^0)$, $t=1,\ldots,T$, are correlated and the matrices C_k^t, $k=1,\ldots,t-1$, $t=1,\ldots,T$, do not have simple forms.

References

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimates. *J. Roy. Statist. Soc. Ser. B* **32**, 283-301.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussions). *J. Roy. Statist. Soc. Ser. B* **36**, 192-236.

Godambe, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika* **63**, 277-84.

Gong, G. and Samaniego, F. (1981). Pseudo maximum likelihood estimation: theory and application. *Ann. Statist.* **9**, 861-869.

Greenless, J. S., Reece, W. S. and Zieschang K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *J. Amer. Statist. Assoc.* **77**, 251-261.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Second edition. Wiley, New York.

Qin, J., Leung, D. and Shao, J. (2002). Estimation with survey data under non-ignorable nonresponse or informative sampling. *J. Amer. Statist. Assoc.* **97**, 193-200.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89**, 846-86.

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

Tang, G., Little, R. J. A. and Raghunathan, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika* **90**, 747-764.

Department of Statistics, University of Wisconsin, Madison, WI 53706, USA.

E-mail: jiang@stat.wisc.edu

School of Finance and Statistics, East China Normal University, Shanghai, 202241, China.

E-mail: shao@stat.wisc.edu

(Received February 2010; accepted August 2011)