Biometrika (2012), *pp.* 1–18
doi: 10.1093/biomet/ass026

Printed in Great Britain

An efficient method of estimation for longitudinal surveys with monotone missing data

BY MING ZHOU AND JAE KWANG KIM

Department of Statistics, Iowa State University, Ames, Iowa 50011, U.S.A.
zhouming@alumni.iastate.edu jkim@iastate.edu

SUMMARY

Panel attrition is frequently encountered in panel sample surveys. When it is related to the observed study variable, the classical approach of nonresponse adjustment using a covariate-dependent dropout mechanism can be biased. We consider an efficient method of estimation with monotone panel attrition when the response probability depends on the previous values of study variable as well as other covariates. Because of the monotone structure of the missing pattern, the response mechanism is missing at random. The proposed estimator is asymptotically optimal in the sense that it minimizes the asymptotic variance of a class of estimators that can be written as a linear combination of the unbiased estimators of the panel estimates for each wave, and incorporates all available information using generalized least squares. Variance estimation is discussed and results from a simulation study are presented.

Some key words: Generalized least squares; Missingness at random; Panel attrition; Propensity score; Survey in time.

1. INTRODUCTION

Longitudinal or panel surveys, in which similar measurements are made on the same sample at different points in time, are very popular in the study of social and physical dynamics that cannot be inferred from cross-sectional surveys. Missing data in the response variable often call into question the validity of statistical analysis, posing in particular a serious challenge to the representativeness of the sample respondents. Bollinger & David (1997, 2001), for example, used the Survey of Income and Program Participation data to show that estimates of food-stamp participation adjusted for nonresponse are significantly different from estimates that do not account for nonresponse.

There is abundant literature on analysing incomplete longitudinal data in the context of biostatistical applications. Diggle et al. (2002), Molenberghs & Kenward (2007) and Fitzmaurice et al. (2008) provide comprehensive overviews of missing data analysis in longitudinal studies. However, most of the literature focuses on likelihood-based approaches under parametric model assumptions and the interest lies in estimating the regression relationship. Such assumptions are rarely used with a complex survey sampling design. In the context of survey sampling, the focus is usually estimation of the population means of the study variables for each wave and the weighting adjustment method is commonly used to handle unit nonresponse. Kalton & Kasprzyk (1986) discussed general issues arising from missing data in longitudinal surveys. Ekholm & Laaksonen (1991), Fuller et al. (1994), Duncan & Stasny (2001), Laaksonen & Chambers (2006), and Laaksonen (2007) have considered nonresponse weighting methods for longitudinal surveys. Rizzo et al. (1996) and Slud & Bailey (2010) evaluated the weighting adjustment method of panel attrition using the Survey of Income and Program Participation data.

The above-mentioned nonresponse adjustment methods for panel surveys presuppose that the implicit response missing mechanism is covariate-dependent, so the response probability depends on the time-invariant baseline covariate X_i but not on the time-variant study variable Y_{it}, according to Little (1995). The nonresponse mechanism is called ignorable if the true response probability depends only on the observed data and not on unobserved random variables. In a panel survey with a monotone missing pattern, ignorable response means that the response probability at time t may depend on X_i and Y_{is} with $s < t$, but not on Y_{it}. The covariate-dependent missing mechanism can be quite restrictive because the dropout mechanism may not be fully explained by demographic baseline covariates. For example, Korinek et al. (2007) analysed the Current Population Survey data using an area level model of response rate on average income and found that the response probability is strongly correlated with household income. They concluded that the current adjustment method, which is essentially based on the covariate-dependent missing assumption, should be rejected.

Likelihood-based approaches, as in Diggle & Kenward (1994), Baker (1995), Little (1995), Molenberghs et al. (1997), and Ibrahim et al. (2001), adopt a parametric approach both for Y_{it} and for the response mechanism. These approaches are not very popular in handling missing data in panel surveys because of the difficulty of the joint modelling and the nonrobustness of the likelihood-based approach in general. Robins et al. (1995) developed a method for the estimation of longitudinal regression models using a semiparametric model for Y_{it} under ignorable response. They assumed a working outcome regression model for Y_{it}, as well as a similar model for the response probability, in developing their estimator. Rotnitzky et al. (1998) extended this method to nonignorable missingness. However, the Robins et al. (1995) method does not make full use of available information, nor does it cover survey sampling. As far as we know, no literature has rigorously discussed parameter estimation in panel surveys when the dropout mechanism may depend on the study variable previously observed.

Under the response model of Robins et al. (1995), we consider an alternative method of parameter estimation that does not assume any model for Y_{it} and uses all available information in the longitudinal data. Under the nonlongitudinal set-up, the proposed estimator reduces to the optimal estimator considered in Cao et al. (2009), which it generalizes for longitudinal surveys. By the orthogonal construction of the control variates, the computation of the optimal estimator is simplified and variance estimation becomes feasible. Furthermore, the proposed method is directly applicable to complex survey sampling.

2. BASIC SET-UP

Let Y_{it} $(i = 1, \ldots, n; t = 0, \ldots, T)$ be the outcome of interest measured on the ith subject at time-point t, and let X_i be auxiliary information that is always observed and remains constant throughout different time-points. We use r_{it} to denote the response indicator for subject i at time-point t: $r_{it} = 1$ if Y_{it} is observed and $r_{it} = 0$ otherwise. Until § 5, we shall regard $(X_i^{\mathrm{T}}, r_{i0}, r_{i1}, \ldots, r_{iT}, Y_{i0}, \ldots, Y_{iT})^{\mathrm{T}}$ as independent and identically distributed random vectors. Assume that the baseline information for subject i, $(X_i^{\mathrm{T}}, Y_{i0})^{\mathrm{T}}$, is always observed. Our goal is to estimate $\mu_t = E(Y_{it})$, the mean of Y_{it}, for $t = 1, \ldots, T$. Let $L_{it} = (X_i^{\mathrm{T}}, Y_{i0}, Y_{i1}, \ldots, Y_{it})^{\mathrm{T}}$ denote the observed values of $(X^{\mathrm{T}}, Y)^{\mathrm{T}}$ for unit i up to time t. For any random variable Δ, we use $\tilde{E}$ to denote the sample average; that is, $\tilde{E}(\Delta) = n^{-1} \sum_{i=1}^{n} \Delta_i$.

Throughout this paper, we shall assume that the missing pattern is monotone, that is,

$$r_{ij} = 0 \Rightarrow r_{i,j+1} = 0 \quad (j = 1, \ldots, T-1). \tag{1}$$

Although the constraint (1) can be somewhat restrictive, monotone missingness covers most realistic situations for panel attrition. The extension to nonmonotone missingness is beyond the scope of this paper. We shall assume the following missing data mechanism:

$$\text{pr}(r_{it} = 1 \mid r_{i,t-1} = 1, L_{iT}) = \text{pr}(r_{it} = 1 \mid r_{i,t-1} = 1, L_{i,t-1}). \tag{2}$$

Equation (2) means that the data are missing at random in the sense of Rubin (1976). See also Little (1995) for its meaning under longitudinal survey set-up. That is to say, at any time-point t, the probability that Y_{it} is missing depends only on what is observed by time $t-1$. In other words, among subjects observed at time $t-1$, the nonresponse probability at time t is unrelated to the current and future outcomes $Y_{it}, \ldots, Y_{iT}$. The missing data mechanism (2) is more realistic than the covariate-dependent missing mechanism, which is often assumed in nonresponse adjustment methods with demographic variables in the response model. In addition to (2), we assume that

$$p_{it} = \text{pr}(r_{it} = 1 \mid r_{i,t-1} = 1, L_{i,t-1}) > \sigma \quad (t = 1, \ldots, T), \tag{3}$$

for some $\sigma > 0$, that is, the response probability for each subject i is bounded away from zero, a necessary mechanism to guarantee the existence of $n^{1/2}$-consistent estimators of μ_t (Robins et al., 1994). The probability p_{it} is the conditional probability of response at time t given the unit i response at time $t-1$. Assumptions (1) and (2) imply that

$$\text{pr}(r_{it} = 1 \mid L_{iT}) = \text{pr}(r_{it} = 1 \mid L_{i,t-1}) = \text{pr}(r_{i0} = 1) \prod_{j=1}^{t} p_{ij}.$$

Written as $\pi_{it} = \pi_{i0} \prod_{j=1}^{t} p_{ij}$, the response probability π_{it} is often called the propensity score (Rosenbaum & Rubin, 1983). When there is no missing subject in the baseline, $\pi_{i0} = 1$. The probability π_{it} is different from p_{it} in that π_{it} refers to the marginal probability of response at time t for subject i, while p_{it} refers to the conditional probability of a response at time t given that unit i responds at time $t-1$. Very often π_{it} depends on $L_{i,t-1}$ and the average of the observed Y_{it}s, thus the following naive estimator

$$\hat{\mu}_{t,\text{naive}} = \frac{\sum_{i=1}^{n} r_{it} Y_{it}}{\sum_{i=1}^{n} r_{it}}, \tag{4}$$

will in general be biased for μ_t. In this case, it is common to model the response probability and use the estimated response probability to obtain the propensity score estimators.

3. Optimal propensity score estimation

3·1. *Propensity score estimation*

For simplicity, let us start from the $T = 1$ case. We now absorb Y_{i0} into X_i, and denote it as X_i. The outcome of interest is then Y_{i1} and we are interested in estimating $\mu_1 = E(Y_{i1})$. Let the true response probability be parametrically modelled by $\pi_{i1} = \pi_1(X_i; \phi_1)$, for some function $\pi_1(\cdot)$ known up to ϕ_1. If the maximum likelihood estimator of ϕ_1, the solution to

$$S_1(\phi_1) = n\tilde{E}\left\{(r_1 - \pi_1)\frac{\partial \pi_1/\partial \phi_1}{\pi_1(1-\pi_1)}\right\} = 0, \tag{5}$$

as denoted by $\hat{\phi}_1$, is available, then the propensity score adjusted estimator of μ_1, denoted by $\hat{Y}_{1,\text{PS}}$, can be computed by solving

$$\hat{U}_{1,\text{PS}} = \tilde{E}\left\{r_1 \hat{\pi}_1^{-1}(Y_1 - \mu_1)\right\} = 0, \tag{6}$$

for μ_1. Inverse probability weighted estimating equations have been previously considered by Horvitz & Thompson (1952), Manski & Lerman (1977), Flanders & Greenland (1991) and Robins et al. (1995) among others. Strictly speaking, the propensity score estimator in (6) is also a function of $\hat{\phi}_1$ that is computed from (5). To discuss the asymptotic variance of the propensity score estimator, we introduce the following lemma.

LEMMA 1. *Suppose $z_1, \ldots, z_n$ are independent and identically distributed random vectors and $\hat{\gamma}$ is the solution to $\tilde{E}\{U(z;\gamma)\} = 0$. Let $U_i(\gamma) = U(z_i;\gamma)$. If* (i) *$E\{U(\gamma^*)\} = 0$ and $\hat{\gamma} = \gamma^* + o_p(1)$, where γ^* is an interior point of the parameter space;* (ii) *var$\{U(\gamma^*)\}$ is finite;* (iii) *$U(\gamma)$ is continuously differentiable in a neighbourhood $\mathcal{N}$ of γ^*;* (iv) *$E\{\partial U(\gamma^*)/\partial\gamma\}$ exists and is nonsingular;* (v) *$E\left\{\sup_{\gamma\in\mathcal{N}} \|\partial U(\gamma)/\partial\gamma\|\right\} < \infty$, then*

$$\hat{\gamma} - \gamma^* = -\left[E\left\{\partial U(\gamma^*)/\partial\gamma^{\mathrm{T}}\right\}\right]^{-1} \tilde{E}\{U(\gamma^*)\} + o_p(n^{-1/2}).$$

This is an immediate application of Lemma 4.3 and Theorem 3.1 of Newey & McFadden (1994).

Remark 1. Consider $U_i(\gamma) = \{\theta^{\mathrm{T}} - g_i(\phi)^{\mathrm{T}}, \psi_i(\phi)^{\mathrm{T}}\}^{\mathrm{T}}$, where $\gamma = (\theta^{\mathrm{T}}, \phi^{\mathrm{T}})^{\mathrm{T}}$ such that θ is distinct from ϕ, $S(\phi) = \sum_i \psi_i(\phi)$ is the score function for ϕ. Let $\hat{\gamma}$ be the solution to $\tilde{E}\{U(\gamma)\} = 0$, that is, $\hat{\gamma} = (\hat{\theta}^{\mathrm{T}}, \hat{\phi}^{\mathrm{T}})^{\mathrm{T}}$, where $\hat{\theta} = \tilde{E}\{g(\hat{\phi})\}$. Then, by Lemma 1, we have

$$\begin{pmatrix} \hat{\theta} - \theta^* \\ \hat{\phi} - \phi^* \end{pmatrix} = -\begin{bmatrix} I & -E\{\partial g(\phi^*)/\partial\phi^{\mathrm{T}}\} \\ 0 & E\{\partial\psi(\phi^*)/\partial\phi^{\mathrm{T}}\} \end{bmatrix}^{-1} \begin{bmatrix} \theta^* - \tilde{E}\{g(\phi^*)\} \\ \tilde{E}\{\psi(\phi^*)\} \end{bmatrix} + o_p(n^{-1/2}).$$

By Pierce (1982), we have $-E\{\partial g(\phi^*)/\partial\phi^{\mathrm{T}}\} = E\{g(\phi^*)\psi(\phi^*)^{\mathrm{T}}\} = \mathrm{cov}\{g(\phi^*), \psi(\phi^*)\}$ and $-E\{\partial\psi(\phi^*)/\partial\phi^{\mathrm{T}}\} = E\{\psi(\phi^*)\psi(\phi^*)^{\mathrm{T}}\} = \mathrm{var}\{\psi(\phi^*)\}$. Therefore, $\hat{\theta}$ can be expressed as

$$\hat{\theta} = \tilde{E}\{g(\phi^*)\} - \mathrm{cov}\{g(\phi^*), \psi(\phi^*)\}\mathrm{var}\{\psi(\phi^*)\}^{-1}\tilde{E}\{\psi(\phi^*)\} + o_p(n^{-1/2}).$$

This implies that

$$\mathrm{var}(\hat{\theta}) = \mathrm{var}[\tilde{E}\{g(\hat{\phi})\}] = \mathrm{var}[\tilde{E}\{g(\phi^*)\} \mid S^{\perp}] + o(n^{-1}), \tag{7}$$

where $\mathrm{var}\{\tilde{E}\{g(\phi^*) \mid S^{\perp}\} = \mathrm{var}[\tilde{E}\{g(\phi^*)\}] - \mathrm{cov}[\tilde{E}\{g(\phi^*)\}, S]\mathrm{var}(S)^{-1}\mathrm{cov}[S, \tilde{E}\{g(\phi^*)\}]$.

By (7),

$$\mathrm{var}\left\{\tilde{E}\left(\frac{r_1 Y_1}{\hat{\pi}_1}\right)\right\} \approx \mathrm{var}\left\{\tilde{E}\left(\frac{r_1 Y_1}{\pi_1}\right) \mid S_1^{\perp}\right\} \leqslant \mathrm{var}\left\{\tilde{E}\left(\frac{r_1 Y_1}{\pi_1}\right)\right\}.$$

Such contradictory phenomena have been discussed by Rosenbaum (1987), Robins et al. (1994), Little & Vartivarian (2005), Kim & Kim (2007) and Beaumont & Bocci (2009). See also Henmi & Eguchi (2004).

3·2. *Optimal propensity score estimation*

We now discuss optimal propensity score estimation. We assume that the propensity score is computed as in (5). In general, the propensity score estimator $\hat{X}_{\mathrm{PS}}$ applied to $\mu_X = E(X)$ is not equal to the complete sample estimator $\hat{X}_n = \tilde{E}(X)$. Thus, the latter estimator $\hat{X}_n$ can be used to improve the efficiency of the propensity score estimator. To combine all the available

information, we consider minimizing

$$Q = \begin{pmatrix} \hat{X}_0 - \mu_X \\ \hat{X}_1 - \mu_X \\ \hat{Y}_1 - \mu_1 \end{pmatrix}^{\mathrm{T}} \mathrm{var} \begin{pmatrix} \hat{X}_0 \\ \hat{X}_1 \\ \hat{Y}_1 \end{pmatrix}^{-1} \begin{pmatrix} \hat{X}_0 - \mu_X \\ \hat{X}_1 - \mu_X \\ \hat{Y}_1 - \mu_1 \end{pmatrix} \tag{8}$$

with respect to μ_X and μ_1, where $\hat{X}_0$ and $\hat{X}_1$ are two unbiased estimators of μ_X and $\hat{Y}_1$ is an unbiased estimator of μ_1. The estimator obtained from minimizing Q in (8) is essentially the generalized least squares estimator. Under the missing data set-up where X_i is always observed and Y_{i1} is subject to missingness, if we know π_{i1}, then we can evaluate $\hat{X}_0 = \tilde{E}(X)$, $\hat{X}_1 = \tilde{E}(r_1 X/\pi_1)$ and $\hat{Y}_1 = \tilde{E}(r_1 Y_1/\pi_1)$. In this case, the estimator that minimizes (8) is given by

$$\hat{\mu}_1 = \tilde{E}\{r_1 Y_1/\pi_1 - (r_1/\pi_1 - 1)X^{\mathrm{T}} B^*\} = \hat{Y}_1 - (\hat{X}_1 - \hat{X}_0)^{\mathrm{T}} B^*,$$

where $B^* = E\{(1/\pi_1 - 1)XX^{\mathrm{T}}\}^{-1} E\{(1/\pi_1 - 1)XY_1\}$. In practice, we can estimate B^* by

$$\hat{\mu}_{1,\mathrm{opt}} = \tilde{E}\{r_1 Y_1/\pi_1 - (r_1/\pi_1 - 1)X^{\mathrm{T}} \hat{B}^*\}, \tag{9}$$

where $\hat{B}^* = \tilde{E}\{r_1 \pi_1^{-1}(\pi_1^{-1} - 1)XX^{\mathrm{T}}\}^{-1} \tilde{E}\{r_1 \pi_1^{-1}(\pi_1^{-1} - 1)XY_1\}$.

The estimator in (9) is asymptotically optimal among the class of linear functions of $\hat{X}_0$, $\hat{X}_1$ and $\hat{Y}_1$ that are unbiased for μ_1. If the true propensity scores are unknown, we can use $\hat{X}_0 = \tilde{E}(X)$, $\hat{X}_1 = \hat{X}_{\mathrm{PS}} = \tilde{E}(r_1 X/\hat{\pi}_1)$ and $\hat{Y}_1 = \hat{Y}_{1,\mathrm{PS}} = \tilde{E}(r_1 Y_1/\hat{\pi}_1)$, where $\hat{\pi}_1 = \pi_1(X; \hat{\phi}_1)$, with $\hat{\phi}_1$ the maximum likelihood estimator given by (5). In this case, the optimal estimator of μ_X still equals $\tilde{E}(X)$, but the optimal estimator of $\hat{\mu}_{1,\mathrm{opt}}$ in (9) using $\hat{\pi}_{i1}$ instead of π_{i1} is not optimal because the covariance between $\hat{Y}_{1,\mathrm{PS}}$ and $(\hat{X}_{\mathrm{PS}}, \hat{X}_n)$ is different from that between $\tilde{E}(r_1 Y_1/\pi_1)$ and $(\tilde{E}(r_1 X/\pi_1), \hat{X}_n)$. To construct an optimal estimator, we can consider one of the form

$$\hat{\mu}_{1,B} = \hat{Y}_{1,\mathrm{PS}} - (\hat{X}_{\mathrm{PS}} - \hat{X}_n)^{\mathrm{T}} B$$

indexed by B and find the quantity B^* that minimizes the variance of $\hat{\mu}_{1,B}$, i.e.,

$$B^* = \left\{\mathrm{var}(\hat{X}_{\mathrm{PS}} - \hat{X}_n)\right\}^{-1} \mathrm{cov}(\hat{X}_{\mathrm{PS}} - \hat{X}_n, \hat{Y}_{1,\mathrm{PS}}).$$

As $E(r_1 X/\pi_1 - X) = 0$ and $E(r_1 X/\pi_1) = \mu_1$, by (7), we can approximate B^* by $\mathrm{var}(\zeta_1 \mid S_1^{\perp})^{-1}\mathrm{cov}(\zeta_1, \zeta_2 \mid S_1^{\perp})$, where $\zeta_1 = \tilde{E}(r_1 X/\pi_1) - \hat{X}_n$, $\zeta_2 = \tilde{E}(r_1 Y_1/\pi_1)$ and $\mathrm{cov}(\zeta_1, \zeta_2 \mid S_1^{\perp}) = \mathrm{cov}(\zeta_1, \zeta_2) - \mathrm{cov}(\zeta_1, S_1)\mathrm{var}(S_1)^{-1}\mathrm{cov}(S_1, \zeta_2)$. Thus, ignoring the smaller order terms in (7), the optimal estimator in (8) with $\hat{X}_0 = \hat{X}_n$, $\hat{X}_1 = \hat{X}_{\mathrm{PS}}$, $\hat{Y}_1 = \hat{Y}_{1,\mathrm{PS}}$ can be obtained by minimizing

$$Q = (\hat{Z} - \mu_Z)^{\mathrm{T}} \mathrm{var}(\hat{Z}_0 \mid S_1^{\perp})^{-1} (\hat{Z} - \mu_Z), \tag{10}$$

where $\hat{Z} = (\hat{X}_n^{\mathrm{T}}, \hat{X}_{\mathrm{PS}}^{\mathrm{T}}, \hat{Y}_{1,\mathrm{PS}})^{\mathrm{T}}$, $\hat{Z}_0 = (\hat{X}_n^{\mathrm{T}}, \tilde{E}(r_1 X^{\mathrm{T}}/\pi_1), \tilde{E}(r_1 Y_1/\pi_1))^{\mathrm{T}}$ and $\mu_Z = (\mu_X^{\mathrm{T}}, \mu_X^{\mathrm{T}}, \mu_1)^{\mathrm{T}}$. The optimal Q in (10) can be also obtained by minimizing

$$Q^* = \begin{pmatrix} \hat{Z} - \mu_Z \\ S_1 \end{pmatrix}^{\mathrm{T}} \mathrm{var} \begin{pmatrix} \hat{Z}_0 \\ S_1 \end{pmatrix}^{-1} \begin{pmatrix} \hat{Z} - \mu_Z \\ S_1 \end{pmatrix}, \tag{11}$$

where $S_1(\phi_1)$ is the score function of ϕ_1, defined in (5). To see this, note that Q^* can be written as $Q^*(\phi_1, \mu_z) = Q_1(\phi_1, \mu_z) + Q_2(\phi_1)$, where $Q_1(\phi_1, \mu_z) = \{\hat{Z} - \mathrm{cov}(\hat{Z}_0, S_1)\mathrm{var}(S_1)^{-1} S_1 - \mu_Z\}^{\mathrm{T}} \mathrm{var}(\hat{Z}_0 \mid S_1^{\perp})^{-1} \{\hat{Z} - \mathrm{cov}(\hat{Z}_0, S_1)\mathrm{var}(S_1)^{-1} S_1 - \mu_Z\}$ and $Q_2(\phi_1) = S_1^{\mathrm{T}} \mathrm{var}(S_1)^{-1} S_1$. Since

$S(\hat{\phi}_1) = 0$, we have $Q_2(\hat{\phi}_1) = 0$ and $Q^*(\hat{\phi}_1, \mu_z) = Q_1(\hat{\phi}_1, \mu_z)$, which is equal to (10). Because the minimum value of $Q_2(\phi_1)$ is zero, the optimal estimator of μ_Z for (10) can also be computed by minimizing (11), when the score equation is used to estimate ϕ_1. Thus, the effect of using the estimated propensity score can be easily taken into account by simply adding the score function for ϕ_1 into the Q term. Furthermore, as discussed in Theorem 2, the inclusion of the score function into the Q term facilitates the linearization for variance estimation. The following example gives an explicit formula for the optimal estimator when ϕ_1 is estimated by its maximum likelihood estimator.

Example 1. Under the response model where the score function for ϕ_1 is

$$S_1(\phi_1) = n\tilde{E}\left[\{r_1 - \pi_1(\phi_1)\}\, h(\phi_1)\right],$$

the coefficient B^* corresponding to the optimal estimators in the family

$$\hat{\mu}_{1,B} = \hat{Y}_{1,\mathrm{PS}} - (\hat{X}_{\mathrm{PS}} - \hat{X}_n)^{\mathrm{T}} B,$$

is $B^* = (V_{XX} - V_{XS}V_{SS}^{-1}V_{SX})^{-1}(V_{XY} - V_{XS}V_{SS}^{-1}V_{SY})$, where

$$\begin{pmatrix} V_{XX} & V_{XY} & V_{XS} \\ V_{YX} & V_{YY} & V_{YS} \\ V_{SX} & V_{SY} & V_{SS} \end{pmatrix} = \mathrm{var}\begin{bmatrix} \tilde{E}\{(r_1/\pi_1 - 1)X\} \\ \tilde{E}\{r_1 Y_1/\pi_1\} \\ \tilde{E}\{(r_1/\pi_1 - 1)\pi_1 h\} \end{bmatrix}.$$

Thus, a consistent estimator of B^* is

$$\hat{B}^* = (I_q, 0)\tilde{E}\left\{\frac{r_1}{\hat{\pi}_1}\left(\frac{1}{\hat{\pi}_1} - 1\right)\begin{pmatrix} X \\ \hat{\pi}_1\hat{h} \end{pmatrix}\begin{pmatrix} X \\ \hat{\pi}_1\hat{h} \end{pmatrix}^{\mathrm{T}}\right\}^{-1} \tilde{E}\left\{\frac{r_1}{\hat{\pi}_1}\left(\frac{1}{\hat{\pi}_1} - 1\right)\begin{pmatrix} X \\ \hat{\pi}_1\hat{h} \end{pmatrix} Y_1\right\},$$

where $q = \dim(X)$, and the resulting optimal estimator is

$$\hat{Y}_{1,\mathrm{opt}} = \hat{Y}_{1,\mathrm{PS}} - (\hat{X}_{\mathrm{PS}} - \hat{X}_n)^{\mathrm{T}}\hat{B}^*. \tag{12}$$

The estimator in (12) is equal to the optimal estimator presented in Cao et al. (2009) in the context of a doubly robust estimator. A similar but slightly different approach was proposed by Tan (2006). However, our derivation of the optimal estimator in (12) is different from those of Cao et al. (2009) and Tan (2006). In addition, the optimal estimator in (12) can be easily generalized to missing data in longitudinal surveys, which will be discussed in the next section.

4. PROPOSED METHOD FOR INCOMPLETE LONGITUDINAL DATA

The proposed optimal estimator in § 3 is based on generalized least squares and can easily be extended to optimal estimation with missing data in longitudinal surveys. To correctly account for the ignorable dropout mechanism in (2), we shall assume a parametric model $p_{it}(L_{i,t-1}; \phi_t)$ for p_{it} in (3). We make no explicit assumptions for the marginal distribution of $L_{i,T}$. We use only the response model and assume no parametric model for Y_{it}. The partial likelihood for ϕ_t is

$$\mathcal{L}(\phi_1, \ldots, \phi_T) = \prod_{t=1}^{T}\prod_{i=1}^{n}\left\{p_{it}^{r_{it}}(1 - p_{it})^{1-r_{it}}\right\}^{r_{i,t-1}} = \prod_{t=1}^{T}\mathcal{L}_t(\phi_t),$$

and the score function for ϕ_t is $S_t(\phi_t) = \partial \log \mathcal{L}_t(\phi_t)/\partial\phi_t$. Writing $h_t(L_{t-1}; \phi_t) = \partial \text{logit}(p_t)/\partial\phi_t$ where $\text{logit}(p) = \log\{p/(1-p)\}$, the score function reduces to $S_t(\phi_t) = n\tilde{E}\left\{r_{t-1}(r_t - p_t)h_t(L_{t-1})\right\}$. When p_t follows the logistic regression model $p_t = \left\{1 + \exp(-\phi_t^{\mathrm{T}} L_{t-1})\right\}^{-1}$, we have $h_t(L_{t-1}; \phi_t) = L_{t-1}$.

At each time-point t, we can construct propensity score estimators $\tilde{E}(r_t X/\hat{\pi}_t)$, $\tilde{E}(r_t Y_1/\hat{\pi}_t), \ldots, \tilde{E}(r_t Y_t/\hat{\pi}_t)$ for $\mu_X, \mu_1, \ldots, \mu_t$, respectively. To incorporate all available information, we use the generalized least squares method in § 3. Denote $u_t = r_t/p_t - 1$ $(t = 1, \ldots, T)$, and

$$\xi_{t-1} = \left(r_0 u_1 \pi_0^{-1} L_0^{\mathrm{T}}, r_1 u_2 \pi_1^{-1} L_1^{\mathrm{T}}, \ldots, r_{t-1} u_t \pi_{t-1}^{-1} L_{t-1}^{\mathrm{T}}\right)^{\mathrm{T}}. \tag{13}$$

where $L_0 = X$ and $L_j = (X^{\mathrm{T}}, Y_1, \ldots, Y_j)^{\mathrm{T}}$ for $j = 1, \ldots, t-1$. Here, Y_0 is absorbed into X. By the definition of u_t, we have $E(u_t \mid L_{t-1}) = 0$ and $E(r_{t-1} u_t \pi_{t-1}^{-1} L_{t-1}) = E\{r_{t-1}\pi_{t-1}^{-1} L_{t-1} E(u_t \mid L_{t-1}, r_1 = \cdots = r_{t-1} = 1)\} = 0$. Thus, $E(\xi_{t-1}) = 0$. By the definition of the response probabilities, $r_{t-1}\hat{u}_t/\hat{\pi}_{t-1} = r_t/\hat{\pi}_t - r_{t-1}/\hat{\pi}_{t-1}$, $\hat{u}_t = u_t(\hat{\phi}_t)$ and $\tilde{E}(r_{t-1}\hat{u}_t\hat{\pi}_{t-1}^{-1} L_{t-1})$ is the difference between the propensity score estimators of $\mu_X, \mu_1, \ldots, \mu_{t-1}$ available at time-point t and those available at time-point $t-1$. Thus, $\tilde{E}(\hat{\xi}_{t-1})$ makes use of all available information up to time-point $t-1$. Also, to incorporate the score functions of $\phi_1, \ldots, \phi_t$ into the generalized least squares, consider

$$\psi_{t-1} = \left(r_0 u_1 p_1 h_1^{\mathrm{T}}, r_1 u_2 p_2 h_2^{\mathrm{T}}, \ldots, r_{t-1} u_t p_t h_t^{\mathrm{T}}\right)^{\mathrm{T}}, \tag{14}$$

where $h_t = h_t(L_{t-1}; \phi_t)$. Note that $n\tilde{E}(r_{t-1} u_t p_t h_t)$ is the score function for ϕ_t. For each time-point t, $E(r_{t-1} u_t p_t h_t) = 0$, because $E(r_{t-1} u_t p_t h_t \mid L_{t-1}, r_1 = \cdots = r_{t-1} = 1) = r_{t-1} p_t h_t E(u_t \mid L_{t-1}, r_1 = \cdots = r_{t-1} = 1) = 0$. Thus, $E(\psi_{t-1}) = 0$. Therefore, using the fact that $E(\psi_{t-1}) = 0$, $E(\xi_{t-1}) = 0$, we estimate $E(Y_t)$ by minimizing

$$Q_t = \begin{pmatrix} \tilde{E}(r_t Y_t/\hat{\pi}_t) - \mu_t \\ \tilde{E}(\hat{\xi}_{t-1}) \\ \tilde{E}(\psi_{t-1}) \end{pmatrix}^{\mathrm{T}} \text{var} \begin{pmatrix} \tilde{E}(r_t Y_t/\pi_t) \\ \tilde{E}(\xi_{t-1}) \\ \tilde{E}(\psi_{t-1}) \end{pmatrix}^{-1} \begin{pmatrix} \tilde{E}(r_t Y_t/\hat{\pi}_t) - \mu_t \\ \tilde{E}(\hat{\xi}_{t-1}) \\ \tilde{E}(\psi_{t-1}) \end{pmatrix}, \tag{15}$$

where $\hat{\xi}_{t-1}$ is ξ_{t-1} after inserting the maximum likelihood estimator $\hat{\phi}_1, \ldots, \hat{\phi}_t$. The control variate $\hat{\xi}_{t-1}$ is included to incorporate all available information up to time-point $t-1$ and the control variate $\tilde{E}(\psi_{t-1})$ is included to incorporate the score equation for $(\phi_1^{\mathrm{T}}, \ldots, \phi_{t-1}^{\mathrm{T}})^{\mathrm{T}}$. For $t = 1$, $\tilde{E}(\hat{\xi}_0) = \hat{X}_{\mathrm{PS}} - \hat{X}_n$ and $\tilde{E}(\psi_0) = n^{-1} S_1$, as discussed in § 3·2. We can write $(S_1^{\mathrm{T}}, \ldots, S_T^{\mathrm{T}})^{\mathrm{T}} = n\tilde{E}(\psi_{T-1})$.

The following theorem gives a form to our optimal estimator for μ_t. Owing to the orthogonality of $r_0 u_1, \ldots, r_{t-1} u_t$, the t subvectors of $\tilde{E}(\xi_{t-1})$ and $\tilde{E}(\psi_{t-1})$ are also orthogonal and $\text{var}\{\tilde{E}(\xi_{t-1})\}$, $\text{var}\{\tilde{E}(\psi_{t-1})\}$, $\text{cov}\{\tilde{E}(\xi_{t-1}), \tilde{E}(\psi_{t-1})\}$ are all block diagonal matrices. This orthogonality of the control variates simplifies the computation of the resulting optimal estimator.

THEOREM 1. *Under the regularity conditions given in the Appendix and the response model such that the score equation for* $(\phi_1^{\mathrm{T}}, \ldots, \phi_T^{\mathrm{T}})^{\mathrm{T}}$ *is* $\tilde{E}(\psi_{T-1}) = 0$, *the optimal value of* $\mu_t = E(Y_t)$ *in the sense of minimizing* (15) *is* $\mu_t^* = E(r_t Y_t/\hat{\pi}_t) - B_t^{*\mathrm{T}} \tilde{E}(\hat{\xi}_{t-1})$, *where* $B_t^* = (B_{1t}^{*\mathrm{T}}, \ldots, B_{tt}^{*\mathrm{T}})^{\mathrm{T}}$, $B_{jt}^* = (I_{dim(L_{j-1})}, 0) D_{jt}^*$ *and*

$$D_{jt}^* = E\left\{(p_j^{-1} - 1)\begin{pmatrix} \pi_{j-1}^{-1} L_{j-1} \\ p_j h_j \end{pmatrix}\begin{pmatrix} L_{j-1} \\ \pi_j h_j \end{pmatrix}^{\mathrm{T}}\right\}^{-1} E\left\{(p_j^{-1} - 1)\begin{pmatrix} \pi_{j-1}^{-1} L_{j-1} \\ p_j h_j \end{pmatrix}\frac{r_t Y_t}{\pi_t}\right\}. \tag{16}$$

The resulting optimal estimator, which is asymptotically equivalent to μ_t^*, is

$$\hat{Y}_{t,opt} = \tilde{E}(r_t Y_t/\hat{\pi}_t) - \sum_{j=1}^{t} \hat{B}_{jt}^{\mathrm{T}} \tilde{E}\left(r_{j-1}\hat{u}_j L_{j-1}/\hat{\pi}_{j-1}\right), \tag{17}$$

where $\hat{B}_{jt} = (I_{dim(L_{j-1})}, 0)\hat{D}_{jt}$,

$$\hat{D}_{jt} = \tilde{E}\left\{\frac{r_t}{\hat{\pi}_t}\hat{u}_j \begin{pmatrix} \hat{\pi}_{j-1}^{-1} L_{j-1} \\ \hat{p}_j \hat{h}_j \end{pmatrix} \begin{pmatrix} L_{j-1} \\ \hat{\pi}_j \hat{h}_j \end{pmatrix}^{\mathrm{T}}\right\}^{-1} \tilde{E}\left\{\frac{r_t}{\hat{\pi}_t}\hat{u}_j \begin{pmatrix} \hat{\pi}_{j-1}^{-1} L_{j-1} \\ \hat{p}_j \hat{h}_j \end{pmatrix} Y_t\right\},$$

$\hat{u}_j = r_j/\hat{p}_j - 1$, $\hat{p}_j = p_j(L_{j-1}; \hat{\phi}_j)$ *and* $\hat{h}_j = h_j(L_{j-1}; \hat{\phi}_j)$.

Proof. See the Appendix. □

By the argument for § 3·2, the optimal estimator in (17) minimizes the asymptotic variance among the class of linear estimators of all available estimators up to time-point t. At time-point t, we have $t+1$ unbiased estimators of μ_X and $t-s+1$ unbiased estimators of μ_s, for $s \leqslant t$.

Remark 2. For $t = 1$, $r_0 \equiv 1$, $\pi_0 \equiv 1$, the optimal estimator in (17) is

$$\tilde{E}(\hat{r}_1 Y_1/\hat{\pi}_1) - \hat{B}_{1,1}^{\mathrm{T}} \tilde{E}\left(\frac{r_0}{\pi_0}\hat{u}_1 L_0\right) = \tilde{E}\left\{\frac{r_1}{\hat{p}_1} Y_1 - \hat{B}_{1,1}^{\mathrm{T}}\left(\frac{r_1}{\hat{p}_1} - 1\right) X\right\},$$

where

$$\hat{B}_{1,1} = (I_{\mathrm{dim(X)}}, 0)\tilde{E}\left\{\frac{r_1}{\hat{p}_1}\left(\frac{1}{\hat{p}_1} - 1\right)\begin{pmatrix} X \\ \hat{p}_1\hat{h} \end{pmatrix}\begin{pmatrix} X \\ \hat{p}_1\hat{h} \end{pmatrix}^{\mathrm{T}}\right\}^{-1} \tilde{E}\left\{\frac{r_1}{\hat{p}_1}\left(\frac{1}{\hat{p}_1} - 1\right)\begin{pmatrix} X \\ \hat{p}_1\hat{h} \end{pmatrix} Y_1\right\},$$

which is the estimator (12) given in Example 1.

We now discuss some asymptotic properties of the optimal estimator in (17). Strictly speaking, $\hat{Y}_{t,\mathrm{opt}}$ is a function of $(\hat{\phi}_1, \ldots, \hat{\phi}_t)$ and should be written as $\hat{Y}_{t,\mathrm{opt}}(\hat{\phi}_1, \ldots, \hat{\phi}_t)$. We show in Theorem 2 that we can safely ignore the effects of $\hat{\phi}_1, \ldots, \hat{\phi}_t$ in $\hat{Y}_{t,\mathrm{opt}}$. That is, $\hat{Y}_{t,\mathrm{opt}}(\hat{\phi}_1, \ldots, \hat{\phi}_t) = \hat{Y}_{t,\mathrm{opt}}(\phi_1^*, \ldots, \phi_t^*) + o_p(n^{-1/2})$, which is often referred to as the Randles (1982) condition. See Kim & Rao (2009), for further discussion of this condition for variance estimation.

THEOREM 2. *Under the regularity conditions in the Appendix, $\hat{Y}_{t,opt}$ in (17) is asymptotically linear with influence function η_t, where*

$$\eta_t = \frac{r_t Y_t}{\pi_t} - \sum_{j=1}^{t} r_{j-1} u_j D_{jt}^{*\,\mathrm{T}} \begin{pmatrix} \pi_{j-1}^{-1} L_{j-1} \\ p_j h_j \end{pmatrix}, \tag{18}$$

and D_{jt}^* is defined in (16). Thus, as $n \to \infty$,

$$n^{1/2}(\hat{Y}_{t,opt} - \mu_t) \to N\{0, var(\eta_t)\} \tag{19}$$

in distribution and

$$\hat{V}^{-1/2}(\hat{Y}_{t,opt} - \mu_t) \to N(0, 1) \tag{20}$$

in distribution, where $\hat{V} = (n-1)^{-1}\tilde{E}\{\hat{\eta}_t - \tilde{E}(\hat{\eta}_t)\}^2$ *and* $\hat{\eta}_t$ *is* η_t *in* (18) *with the estimated parameters inserted.*

Proof. See the Appendix. □

Remark 3. We obtain $\hat{\mu}_{t,\mathrm{opt}}$ by minimizing Q_t in (15) with respect to μ_t. One may consider estimating $\mu_1, \ldots, \mu_T$ simultaneously by minimizing

$$\tilde{Q}_T = \begin{pmatrix} \tilde{E}(X) - \mu_X \\ \tilde{E}(r_1 Y_1/\hat{\pi}_1) - \mu_1 \\ \vdots \\ \tilde{E}(r_T Y_T/\hat{\pi}_T) - \mu_T \\ \tilde{E}(\hat{\xi}_{T-1}) - E(\xi_{T-1}) \\ \tilde{E}(\psi_{T-1}) - E(\psi_{T-1}) \end{pmatrix}^{\mathrm{T}} \mathrm{var} \begin{pmatrix} \tilde{E}(X) \\ \tilde{E}(r_1 Y_1/\pi_1) \\ \vdots \\ \tilde{E}(r_T Y_T/\pi_T) \\ \tilde{E}(\xi_{T-1}) \\ \tilde{E}(\psi_{T-1}) \end{pmatrix}^{-1} \begin{pmatrix} \tilde{E}(X) - \mu_X \\ \tilde{E}(r_1 Y_1/\hat{\pi}_1) - \mu_1 \\ \vdots \\ \tilde{E}(r_T Y_T/\hat{\pi}_T) - \mu_T \\ \tilde{E}(\hat{\xi}_{T-1}) - E(\xi_{T-1}) \\ \tilde{E}(\psi_{T-1}) - E(\psi_{T-1}) \end{pmatrix}, \quad (21)$$

with respect to $(\mu_X^{\mathrm{T}}, \mu_1, \ldots, \mu_T)^{\mathrm{T}}$. It can be shown that under monotone missingness, minimizing $\tilde{Q}_T$ to estimate $\mu_1, \ldots, \mu_T$ simultaneously is equivalent to minimizing Q_t in (15) for each μ_t. The dimension of the vector in (21) is $2qT + T^2 + q$, which is much larger than the dimension $2qt + t^2 - t + 1$ associated with Q_t in (15), where $q = \dim(X)$.

5. Extension to complex survey sampling

In this section, we extend the result in the previous section to the situation where the sample is selected by a complex sampling design from a finite population indexed by $U_N = \{1, \ldots, N\}$ with known population size N. Let $\mathcal{F}_N = \{(X_i^{\mathrm{T}}, Y_{i1}, \ldots, Y_{iT})^{\mathrm{T}} : i = 1, \ldots, N\}$. We shall assume monotone missingness as described in (1), and adopt a missing-at-random mechanism as in (2). Let A denote the set of indices of the sample elements in the baseline, with fixed sample size n and design weights w_i $(i = 1, \ldots, N)$. Assume that there are no missing data in the population at the baseline, that is, $r_{i0} = 1$. We use A_t to denote the set of sample respondents in time-point t. That is, $A_t = \{i \in U_N : r_{it} = 1\}$. We shall denote $\tilde{E}_{\mathcal{F}}(\Delta) = N^{-1}\sum_{i=1}^N \Delta_i$ and $\tilde{E}_A(\Delta) = N^{-1}\sum_{i\in A} w_i \Delta_i$.

The parameters of interest are the population means of the study variables $\mu_t = N^{-1}\sum_{i=1}^N Y_{it} (t = 1, \ldots, T)$. The score function for ϕ_t is

$$S_t(\phi_t) = n\tilde{E}_A\{r_{t-1}(r_t - p_t)h_t/w\} = \sum_{i\in A} r_{i,t-1}(r_{it} - p_{it})h_{it},$$

where $h_{it} = \partial \mathrm{logit}(p_{it})/\partial\phi_t$. The propensity score estimator for μ_t is

$$\hat{Y}_{t,\mathrm{PS}} = \tilde{E}_A\left(\frac{r_t Y_t}{\hat{\pi}_t}\right) = N^{-1}\sum_{i\in A} w_i \frac{r_{it} Y_{it}}{\hat{\pi}_{it}}.$$

To compute the optimal estimator, we shall adopt ξ_{t-1} in (13), ψ_{t-1} in (14), and construct a Q_t term similar to (15). Note that $E\{\tilde{E}_A(r_t Y_t/\pi_t) \mid \mathcal{F}_N\} = E[\tilde{E}_A\{E(r_t Y_t/\pi_t \mid A, \mathcal{F}_N)\} \mid \mathcal{F}_N] = E\{\tilde{E}_A(Y_t) \mid \mathcal{F}_N\} = \mu_t$. Since $E(u_t \mid A_{t-1}) = 0$, we have $E\{\tilde{E}_A(r_{t-1}\pi_{t-1}^{-1}u_t L_{t-1}) \mid \mathcal{F}_N\} = 0$ and $E\{\tilde{E}_A(\xi_{t-1}) \mid \mathcal{F}_N\} = 0$. Similarly, $E\{\tilde{E}_A(\psi_{t-1}/w) \mid \mathcal{F}_N\} = 0$. Thus we can consider the Q_t term

similar to (15) as

$$Q_t = \begin{pmatrix} \tilde{E}_A(r_t Y_t/\hat{\pi}_t) - \mu_t \\ \tilde{E}_A(\hat{\xi}_{t-1}) \\ \tilde{E}_A(\psi_{t-1}/w) \end{pmatrix}^{\mathrm{T}} \mathrm{var}\left\{ \begin{pmatrix} \tilde{E}_A(r_t Y_t/\pi_t) \\ \tilde{E}_A(\xi_{t-1}) \\ \tilde{E}_A(\psi_{t-1}/w) \end{pmatrix} \middle| \mathcal{F}_N \right\}^{-1} \begin{pmatrix} \tilde{E}_A(r_t Y_t/\hat{\pi}_t) - \mu_t \\ \tilde{E}_A(\hat{\xi}_{t-1}) \\ \tilde{E}_A(\psi_{t-1}/w) \end{pmatrix}. \quad (22)$$

To discuss the asymptotic properties of the propensity score estimators in the complex survey, the following conditions are assumed in addition to Conditions A1–A6 stated in the Appendix for Theorem 1.

Condition 1. The design weights are bounded from above and below, that is, $0 < K_l \leqslant nN^{-1}w_i \leqslant K_u < \infty \quad (i = 1, \ldots, N)$, uniformly in n, where K_l and K_u are fixed constants.

Condition 2. The sample moments with design weight converges to the population moments, that is,

$$N^{-1}\sum_{i\in A} w_i v_i = N^{-1}\sum_{i=1}^{N} v_i + O_p(n^{-1/2}),$$

for any v_i with finite second moments.

COROLLARY 1. *Let $\mathcal{F}_N = \{(X_i^{\mathrm{T}}, Y_{i,1}, \ldots, Y_{i,T})^{\mathrm{T}} : i = 1, \ldots N\}$ be a finite population. A probability sample of size n is selected with design weights w_i. Subject to Conditions* A1–A6, 1 *and* 2, *under the monotone missing pattern and response model such that the score equation for $(\phi_1^{\mathrm{T}}, \ldots, \phi_T^{\mathrm{T}})^{\mathrm{T}}$ is $\tilde{E}_A(\psi_{T-1}/w) = 0$, the optimal value of μ_t minimizing* (22) *is $\mu_t^* = \tilde{E}_A(r_t Y_t/\hat{\pi}_t) - B_t^{*\mathrm{T}}\tilde{E}_A(\hat{\xi}_{t-1})$, where $B_t^* = (B_{1t}^{*\mathrm{T}}, \ldots, B_{tt}^{*\mathrm{T}})^{\mathrm{T}}$, $B_{jt}^* = (I_{dim(L_{j-1})}, 0)D_{jt}^*$ and*

$$D_{jt}^* = \tilde{E}_{\mathcal{F}}\left\{ w(p_j^{-1} - 1)\begin{pmatrix} \pi_{j-1}^{-1}L_{j-1} \\ p_j h_j/w \end{pmatrix}\begin{pmatrix} L_{j-1} \\ \pi_j h_j/w \end{pmatrix}^{\mathrm{T}} \right\}^{-1} \tilde{E}_{\mathcal{F}}\left\{ w(p_j^{-1} - 1)\begin{pmatrix} \pi_{j-1}^{-1}L_{j-1} \\ p_j h_j/w \end{pmatrix} Y_t \right\}. \quad (23)$$

Here, D_{jt}^* can be consistently estimated by

$$\hat{D}_{jt} = \tilde{E}_A\left\{ w\frac{r_t}{\hat{\pi}_t}\hat{u}_j\begin{pmatrix} \hat{\pi}_{j-1}^{-1}L_{j-1} \\ \hat{p}_j\hat{h}_j/w \end{pmatrix}\begin{pmatrix} L_{j-1} \\ \hat{\pi}_j\hat{h}_j/w \end{pmatrix}^{\mathrm{T}} \right\}^{-1} \tilde{E}_A\left\{ w\frac{r_t}{\hat{\pi}_t}\hat{u}_j\begin{pmatrix} \hat{\pi}_{j-1}^{-1}L_{j-1} \\ \hat{p}_j\hat{h}_j/w \end{pmatrix} Y_t \right\}. \quad (24)$$

The resulting optimal estimator, which is asymptotically equivalent to μ_t^*, is

$$\hat{Y}_{t,opt} = \tilde{E}_A(r_t Y_t/\hat{\pi}_t) - \sum_{j=1}^{t}\hat{B}_{jt}^{\mathrm{T}}\tilde{E}_A\left(r_{j-1}\hat{u}_j L_{j-1}/\hat{\pi}_{j-1}\right), \quad (25)$$

where $\hat{B}_{jt} = (I_{dim(L_{j-1})}, 0)\hat{D}_{jt}$, $\hat{u}_j = r_j/\hat{p}_j - 1$, $\hat{\pi}_j = \prod_{k=1}^{j}\hat{p}_k$, $\hat{p}_j = p_j(L_{j-1}; \hat{\phi}_j)$ and $\hat{h}_j = h(L_{j-1}; \hat{\phi}_j)$.

To discuss variance estimation, similar to (18), write

$$\eta_t = \frac{r_t Y_t}{\pi_t} - \sum_{j=1}^{t} r_{j-1}u_j D_{jt}^{*\mathrm{T}}\begin{pmatrix} \pi_{j-1}^{-1}L_{j-1} \\ p_j h_j/w \end{pmatrix}, \quad (26)$$

where D_{jt}^* is defined in (23). Let $\hat{\eta}_t$ be the corresponding plug-in estimator of η_t in (26) with $\hat{D}_{jt}$ in (24); then $\hat{Y}_{t,\text{opt}}$ in (25) can be written as $\hat{Y}_{t,opt} = \tilde{E}_A(\hat{\eta}_t)$. Thus, by similar arguments to the proof of Theorem 2, we can establish the Randles (1982) condition for $\tilde{E}_A(\hat{\eta}_t)$ to get $\tilde{E}_A(\hat{\eta}_t) = \tilde{E}_A(\eta_t) + o_p(n^{-1/2})$, and we can apply the standard complete sample method to estimate the variance of $\tilde{E}_A(\eta_t)$, which is asymptotically equivalent to the variance of $\tilde{E}_A(\hat{\eta}_t)$ (Kim & Rao, 2009).

To calculate $\text{var}\{\tilde{E}_A(\eta_t) \mid \mathcal{F}_N\}$, the reverse framework of Fay (1992), Shao & Steel (1999) and Kim & Rao (2009) is used. Specifically, denote $\mathcal{R}_t = \{r_{11}, \ldots, r_{Nt}\}$ and $\bar{\mathcal{R}}_t = \{\mathcal{R}_1, \ldots, \mathcal{R}_t\}$. Then $\text{var}\{\tilde{E}_A(\eta_t) \mid \mathcal{F}_N\}$ can be written as

$$V_1 + V_2 = E[\text{var}\{\tilde{E}_A(\eta_t) \mid \bar{\mathcal{R}}_t, \mathcal{F}_N\} \mid \mathcal{F}_N] + \text{var}[E\{\tilde{E}_A(\eta_t) \mid \bar{\mathcal{R}}_t, \mathcal{F}_N\} \mid \mathcal{F}_N]. \tag{27}$$

For any g with finite second moment, we assume that $N^{-1}\sum_{i\in A}\sum_{i\in A}\Omega_{ij}g_ig_j$ is an unbiased estimator of $\text{var}\{\tilde{E}_A(g) \mid \mathcal{F}_N\}$ by design, where Ω_{ij} depends on the joint inclusion probability. It follows that $\text{var}\{\tilde{E}_A(\eta_t) \mid \bar{\mathcal{R}}_t, \mathcal{F}_N\}$ in (27) can be estimated by

$$\hat{V}_1(\eta) = N^{-2}\sum_{i\in A}\sum_{j\in A}\Omega_{ij}\eta_{it}\eta_{jt}.$$

To show the consistency of $\hat{V}_1$ for V_1 in (27), we assume that finite fourth moments exist for variables stated in Condition A4, $\sum_{i=1}^N |\Omega_{ij}| = O(n^{-1}N)$, and $\text{var}[n\,\text{var}\{\tilde{E}_A(\eta_t) \mid \bar{\mathcal{R}}_t, \mathcal{F}_N\} \mid \mathcal{F}_N] = o_p(1)$. Consequently, $\hat{V}_1(\eta)$ is consistent for V_1 and $\hat{V}_1(\hat{\eta})$ is also consistent for V_1 under some conditions (Kim et al., 2006). The second term V_2 in (27) is $V_2 = \text{var}[E\{\tilde{E}_A(\eta_t) \mid \bar{\mathcal{R}}_t, \mathcal{F}_N\} \mid \mathcal{F}_N] = \text{var}(N^{-1}\sum_{i=1}^N \eta_{it} \mid \mathcal{F}_N) = \text{var}\{N^{-1}\sum_{i=1}^N(\eta_{it} - Y_{it}) \mid \mathcal{F}_N\}$. Next, using $r_{j-1}u_j/\pi_{j-1} = r_j/\pi_j - r_{j-1}/\pi_{j-1}$, we can write

$$\begin{aligned}\eta_{it} - Y_{it} &= \left(\frac{r_{it}}{\pi_{it}} - 1\right) Y_{it} - \sum_{j=1}^{t} r_{i,j-1}u_{ij} D_{jt}^{*\,\mathrm{T}} \begin{pmatrix} L_{i,j-1}/\pi_{i,j-1} \\ h_{ij}p_{ij}/w_i \end{pmatrix} \\ &= \sum_{j=1}^{t} r_{i,j-1}u_{ij}\left\{\frac{Y_{it}}{\pi_{i,j-1}} - D_{jt}^{*\,\mathrm{T}} \begin{pmatrix} L_{i,j-1}/\pi_{i,j-1} \\ h_{ij}p_{ij}/w_i \end{pmatrix}\right\}.\end{aligned}$$

Recall that $E(r_{j-1}u_j \mid \mathcal{F}_N) = 0$ and $E(r_{i-1}u_i r_{j-1}u_j \mid \mathcal{F}_N) = \pi_{j-1}(1/p_j - 1)I(i=j)$, for any i, j. Then,

$$V_2 = N^{-2}\sum_{i=1}^{N}\sum_{j=1}^{t}\frac{1}{\pi_{i,j-1}}\left(\frac{1}{p_{ij}} - 1\right)\left\{Y_{it} - \pi_{i,j-1}D_{jt}^{*\,\mathrm{T}}\begin{pmatrix} L_{i,j-1}/\pi_{i,j-1} \\ h_{ij}p_{ij}/w_i \end{pmatrix}\right\}^2,$$

which can be estimated by

$$\hat{V}_2 = N^{-2}\sum_{i\in A_t} w_i \frac{1}{\hat{\pi}_{it}}\sum_{j=1}^{t}\frac{1}{\hat{\pi}_{i,j-1}}\left(\frac{1}{\hat{p}_{ij}} - 1\right)\left\{Y_{it} - \hat{\pi}_{i,j-1}\hat{D}_{jt}^{\mathrm{T}}\begin{pmatrix} L_{i,j-1}/\hat{\pi}_{i,j-1} \\ \hat{h}_{ij}\hat{p}_{ij}/w_i \end{pmatrix}\right\}^2.$$

Under Condition 2, we have $\hat{V}_2 = V_2 + o_p(N^{-1})$. Therefore, $\hat{V}\{\tilde{E}_A(\hat{\eta}_t)\} = \hat{V}_1 + \hat{V}_2$ is consistent for the variance of $\hat{Y}_{t,\text{opt}}$ in (25).

The order of the first term V_1 is $V_1 = O_p(n^{-1})$, and the order of the second term V_2 is $V_2 = O_p(N^{-1})$. Thus, when the sampling fraction n/N is negligible, that is, $n/N = o(1)$, the second term V_2 can be ignored, and $\hat{V}_1$ becomes a consistent estimator for the total variance.

6. SIMULATION STUDY

To test our theory and to examine the performance of the proposed estimator for finite sample sizes, we performed a simulation study. Two models were considered to generate two sets of independent samples separately. In the first model, denoted model A, we used a linear regression model with serial correlation. In model A, we generated $(x_i, y_{i0}, y_{i1}, y_{i2}, y_{i3})^{\mathrm{T}}$ where

$$Y_0 = X/2 + e_0, \quad Y_t = 1 + X/2 + Y_{t-1} + e_t, \quad t > 1,$$

$X \sim N(0, 1)$, and the e_t are independent and identically distributed as $N(0, 1)$. In the second model, denoted model B, we used a nonlinear regression model with serial correlation. The model is

$$Y_0 = X/3 + Z/3 + e_0, \quad Y_t = 1 + X/3 + Z/3 + e_t, \quad t > 1, \tag{28}$$

where $X \sim N(0, 1)$, $Z = \mathrm{sgn}(\mathrm{X})|\mathrm{X}|^{1/2} + \epsilon$, with sgn the sign function, ϵ and the e_t are independent and identically distributed $N(0, 1)$ random variables. In both models, the missingness indicators r_{it} for $t = 0, 1, 2$ were also independently generated from Bernoulli distributions with

$$\mathrm{pr}(r_t = 1 \mid X, Y_{t-1}, r_{t-1} = 1) = (1 + \exp[-2{\cdot}5 - X + \{Y_{t-1} - (t-1)\}/2])^{-1},$$

and there are no missing data in the baseline. In both models, the true mean of Y_t is $E(Y_t) = t$. The response rates for Y_1, Y_2, Y_3 are 0·90, 0·83, 0·76 for model A and 0·90, 0·82, 0·74 for model B.

The parameters of interest are $\mu_t = E(Y_t)$, for $t = 1, 2, 3$. We computed five estimators for each: $\tilde{E}(Y_t)$, the full sample estimator under no missingness; the naive estimator in (4) using the simple mean of the responding part of the sample; $\tilde{E}(r_t Y_t/\hat{\pi}_t)$, the direct propensity score estimator; $\hat{Y}_{t,\mathrm{opt}}$, our optimal propensity score adjusted estimator in (17). In addition, we considered an estimator proposed by Robins et al. (1995),

$$\hat{Y}_{t,\mathrm{RRZ}} = \tilde{E}(r_t Y_t/\hat{\pi}_t)/\tilde{E}(r_t/\hat{\pi}_t) - \hat{\beta}_{1t}^{\mathrm{T}} \left\{ \tilde{E}(r_t X/\hat{\pi}_t)/\tilde{E}(r_t/\hat{\pi}_t) - \bar{X}_n \right\},$$

where $\hat{\beta}_{1t}$ is the solution to

$$\tilde{E}\left(\frac{r_t}{\hat{\pi}_t} \left[Y_t - \frac{\tilde{E}(r_t Y_t/\hat{\pi}_t)}{\tilde{E}(r_t/\hat{\pi}_t)} - \beta_{1t}^{\mathrm{T}} \left\{ X - \frac{\tilde{E}(r_t X/\hat{\pi}_t)}{\tilde{E}(r_t/\hat{\pi}_t)} \right\} \right] X \right) = 0.$$

We used 10 000 Monte Carlo samples of size $n = 500$. The simulation results in Table 1 show that the naive estimator is severely biased, and the other three propensity score estimators are almost unbiased. The estimator of Robins et al. (1995) is more efficient than the direct propensity score estimator because it uses the auxiliary information in x, though not in optimal way. However, the estimator of Robins et al. (1995) is less efficient than the optimal estimator. At time $t = 3$, the relative efficiency of the proposed estimator over the estimator of Robins et al. (1995) under model B is 167%, greater than 124% under model A, because the linear regression model does not hold in the sample generated by (28).

We also computed a variance estimator of the optimal estimator using the formula in (20). The relative biases of the variance estimator in (20), for $t = 1, 2, 3$, are 0·026, 0·020, and −0·028

Table 1. *Comparison of five estimators when* $n = 500$, $T = 3$, *where the bias and variance values are multiplied by* 1000

		Model A			Model B		
Parameter	Estimator	Bias	Var	StdMSE	Bias	Var	StdMSE
μ_1	Full	−0·4	6·1	100	1·4	7·9	100
	Naive	22·4	6·7	120	23·4	8·9	119
	Direct	−0·6	6·3	105	1·3	8·3	105
	RRZ	−0·6	6·3	105	1·2	8·3	105
	Optimal	−0·6	6·4	105	1·4	8·3	105
μ_2	Full	−1·0	10·5	100	1·5	14·9	100
	Naive	−75·6	12·1	169	−162·4	16·9	291
	Direct	−2·5	12·2	116	0·8	19·7	132
	RRZ	−2·6	11·6	111	0·1	17·5	118
	Optimal	−1·8	11·5	109	1·7	16·1	108
μ_3	Full	−1·5	16·1	100	1·8	23·8	100
	Naive	−302·9	18·6	687	−595·8	27·4	1605
	Direct	−6·3	52·2	325	−15·7	189·2	795
	RRZ	−8·8	23·5	147	−18·8	51·4	217
	Optimal	−2·2	18·9	118	3·0	30·9	130

Var, variance; StdMSE, standardized mean squared error; Full, full sample estimator; Naive, naive estimator defined in (4); Direct, direct propensity score estimator; RRZ, the estimator proposed by Robins et al. (1995); Optimal, the proposed optimal estimator.

respectively for model A. For model B, the relative biases are 0·014, −0·012, and −0·067 respectively. Thus, the simulation results show good finite sample performance of the proposed variance estimator.

7. Future work

The proposed method makes the best use of all asymptotically unbiased estimators available for each wave of the panel survey, and can be directly applied when the baseline sample is a complex probability sample. Variance estimation is a relatively straightforward extension. The theory was developed only for estimating the mean of Y_t but it can be extended to other parameters, by considering a propensity score method applied to an estimating function $U(\beta)$ that satisfies $E\{U(\beta)\} = 0$.

The proposed method requires that the missing pattern be monotone. If it is applied to nonmonotone missing patterns, estimation of response probability at time t can be more complicated because $Y_{i,t-1}$ is not always observed for nonmonotone missing cases. Extension of the proposed method to nonmonotone missing data will be an important topic for future research.

Acknowledgement

We thank two referees, the associate editor, and editor for their very helpful comments. The research of the second author was supported by a Cooperative Agreement between the U.S. Department of Agriculture Natural Resources Conservation Service and Iowa State University.

Appendix

Proof of Theorem 1

Let $h_{it}(\phi_t) = \partial \text{logit}(p_{it})/\partial\phi_t$, where $\text{logit}(p) = \log\{p/(1-p)\}$, and let $H_{it} = (\xi_{i,t-1}^{\mathrm{T}}, \psi_{i,t-1}^{\mathrm{T}})^{\mathrm{T}}$. Throughout the following arguments, unless explicitly pointed out, we shall suppress the notation of true

parameters ϕ_t^* such that all expectations are evaluated at the true parameters. In addition, $\|\cdot\|$ is used to denote the Euclidean norm. We shall assume the following regularity conditions.

Condition A1. The conditional response probabilities are bounded from below uniformly, that is, there exists a positive constant σ such that $p_{it} > \sigma$ for $i = 1, \ldots, n,\ t = 1, \ldots, T$ uniformly.

Condition A2. The solution $\hat{\phi}_t$ to $S_t(\phi_t) = 0$ satisfies $\hat{\phi}_t = \phi_t^* + o_p(1)$ for $t = 1, \ldots, T$.

Condition A3. In a neighbourhood of ϕ_t^*, $p_{it}(\phi_t)$ is twice continuously differentiable for $t = 1, \ldots, T$.

Condition A4. Finite second moments of X, Y_t, $h_t(\phi_t^*)$, $\partial h_t(\phi_t^*)/\partial\phi_t$ exist for $t = 1, \ldots, T$.

Condition A5. For H_{iT}, $\mathrm{var}(H_{iT})$ is nonsingular, $E(\partial\xi_{T-1}/\partial\bar{\phi}_T)$, $E(\partial\psi_{T-1}/\partial\bar{\phi}_T)$ exist and are nonsingular, where $\bar{\phi}_T = (\phi_1^{\mathrm{T}}, \ldots, \phi_T^{\mathrm{T}})^{\mathrm{T}}$.

Condition A6. There exists a neighbourhood $\mathcal{N}_t$ of ϕ_t^* such that $E\{\sup_{\phi_t\in\mathcal{N}_t}\|h_t(\phi_t)\|\} < \infty$, $E\{\sup_{\phi_t\in\mathcal{N}_t}\|h_t(\phi_t)h_t(\phi_t)^{\mathrm{T}}\|\} < \infty$ and $E(\sup_{\phi_t\in\mathcal{N}_t}\|\partial h_t/\partial\phi_t\|) < \infty$ for $t = 1, \ldots, T$.

Proof. Denote $\bar{\phi}_t = (\phi_1^{\mathrm{T}}, \ldots, \phi_t^{\mathrm{T}})^{\mathrm{T}}$ and $\bar{S}_t(\bar{\phi}_t) = (S_1^{\mathrm{T}}, \ldots, S_t^{\mathrm{T}})^{\mathrm{T}}$ for $t = 1, \ldots, T$. The optimal B_t^* minimizing the variance of $\tilde{E}(r_tY_t/\hat{\pi}_t) - B_t^{\mathrm{T}}\tilde{E}(\hat{\xi}_{t-1})$ satisfies $\mathrm{var}\{\tilde{E}(\hat{\xi}_{t-1})\}B_t^* = \mathrm{cov}\{\tilde{E}(\hat{\xi}_{t-1}), \tilde{E}(r_tY_t/\hat{\pi}_t)\}$. Let $U_{it}(\gamma) = (\mu_t - r_{it}Y_{it}/\pi_{it}, \xi_{i,t-1}^{\mathrm{T}}, \psi_{i,t-1}^{\mathrm{T}})^{\mathrm{T}}$, where $\gamma = (\mu_t, \bar{\phi}_t^{\mathrm{T}})^{\mathrm{T}}$. First of all, conditions (i) and (ii) of Lemma 1 hold by Conditions A1, A2 and A4. For example, since $\pi_{it} = \prod_{j=1}^t p_{ij} \geqslant \sigma^t$, $|r_{it}/\pi_{it} - r_{i,t-1}/\pi_{i,t-1}| \leqslant 2/\sigma^t$, $E(r_tY_t^2/\pi_t^2) \leqslant E(Y_t^2)/\sigma^2$, $E\|(r_t/\pi_t - r_{t-1}/\pi_{t-1})L_{t-1}\|^2 \leqslant 2E\|L_{t-1}\|^2/\sigma^2$. Also, Condition A3 implies (iii), and Condition A5 implies (iv). Note that $p_{it}(1-p_{it})h_{it} = \partial p_{it}/\partial\phi_t$, so $E(\sup_{\phi_t\in\mathcal{N}_t}\|\partial p_{it}/\partial\phi_t\|) \leqslant E\{\sup_{\phi_t\in\mathcal{N}_t}\|h_{it}(\phi_t)\|\}/4 < \infty$, $\|\partial\pi_{it}/\partial\phi_k\| = \|\partial p_{ik}/\partial\phi_k\prod_{j\neq k}p_{ij}(\phi_j)\| \leqslant \|\partial p_{ik}/\partial\phi_k\|$. Moreover, $\|\partial\{(r_{it} - r_{i,t-1}p_{it})h_{it}\}/\partial\phi_t\| \leqslant \|h_{it}h_{it}^{\mathrm{T}}\|/4 + 2\|\partial h_{it}/\partial\phi_t\|$. Therefore, Condition A6 implies (v). By similar arguments to the proof of Lemma 1, we have

$$\begin{aligned}
\hat{\bar{\phi}}_t - \bar{\phi}_t^* &= -E\left\{\partial\bar{S}_t(\bar{\phi}_t^*)/\partial\bar{\phi}_t\right\}^{-1}\bar{S}_t(\bar{\phi}_t^*) + o_p(n^{-1/2}),\\
\tilde{E}(\hat{\xi}_{t-1}) &= \tilde{E}\{\xi_{t-1}(\bar{\phi}_t^*)\} - E\left\{\partial\xi_{t-1}(\bar{\phi}_t^*)/\partial\bar{\phi}_t^{\mathrm{T}}\right\}E\left\{\partial\bar{S}_t(\bar{\phi}_t^*)/\partial\bar{\phi}_t\right\}^{-1}\bar{S}_t(\bar{\phi}_t^*) + o_p(n^{-1/2}),\\
\tilde{E}(r_tY_t/\hat{\pi}_t) &= \tilde{E}\{r_tY_t/\pi_t(\bar{\phi}_t^*)\} - E\left[\partial\{r_tY_t/\pi_t(\bar{\phi}_t^*)\}/\partial\bar{\phi}_t^{\mathrm{T}}\right]E\left\{\partial\bar{S}_t(\bar{\phi}_t^*)/\partial\bar{\phi}_t\right\}^{-1}\bar{S}_t(\bar{\phi}_t^*) + o_p(n^{-1/2}).
\end{aligned}$$

Again (Pierce, 1982), we have

$$E\left(\partial\xi_{t-1}/\partial\bar{\phi}_t^{\mathrm{T}}\right) = -\mathrm{cov}(\xi_{t-1}, \bar{S}_t) = -\mathrm{cov}(\xi_{t-1}, \psi_{t-1}),\ E\left(\partial\bar{S}_t/\partial\bar{\phi}_t\right) = \mathrm{var}(\bar{S}_t) = n\,\mathrm{var}(\psi_{t-1}).$$

Therefore,

$$\begin{aligned}
\mathrm{var}\{\tilde{E}(\hat{\xi}_{t-1})\} &= \mathrm{var}\{\tilde{E}(\xi_{t-1})\} - \mathrm{cov}(\xi_{t-1}, \bar{S}_t)\mathrm{var}(\bar{S}_t)^{-1}\mathrm{cov}(\bar{S}_t, \xi_{t-1}) + o(n^{-1}),\\
\mathrm{cov}\{\tilde{E}(\hat{\xi}_{t-1}), \tilde{E}(r_tY_t/\hat{\pi}_t)\} &= \mathrm{cov}\{\tilde{E}(\xi_{t-1}), \tilde{E}(r_tY_t/\pi_t)\}\\
&\quad - \mathrm{cov}(\xi_{t-1}, \bar{S}_t)\mathrm{var}(\bar{S}_t)^{-1}\mathrm{cov}(\bar{S}_t, r_tY_t/\pi_t) + o(n^{-1}).
\end{aligned}$$

Let

$$\mathrm{var}\begin{pmatrix}\xi_{t-1}\\ \psi_{t-1}\end{pmatrix} = E\left\{\begin{pmatrix}\xi_{t-1}\\ \psi_{t-1}\end{pmatrix}\begin{pmatrix}\xi_{t-1}\\ \psi_{t-1}\end{pmatrix}^{\mathrm{T}}\right\} = \begin{pmatrix}V_{LL,t} & V_{LS,t}\\ V_{SL,t} & V_{SS,t}\end{pmatrix},\ E\left\{\begin{pmatrix}\xi_{t-1}\\ \psi_{t-1}\end{pmatrix}\frac{r_tY_t}{\pi_t}\right\} = \begin{pmatrix}V_{LY,t}\\ V_{SY,t}\end{pmatrix}.$$

Then

$$\begin{aligned}
B_t^* &= (V_{LL,t} - V_{LS,t}V_{SS,t}^{-1}V_{SL,t})^{-1}(V_{LY,t} - V_{LS,t}V_{SS,t}^{-1}V_{SY,t}) + o_p(1)\\
&= (I, 0)\begin{pmatrix}V_{LL,t} & V_{LS,t}\\ V_{SL,t} & V_{SS,t}\end{pmatrix}^{-1}\begin{pmatrix}V_{LY,t}\\ V_{SY,t}\end{pmatrix} + o_p(1).
\end{aligned}$$

Since $E(r_{i-1}u_i|L_{i-1})=0$, $E(r_{i-1}u_i^2|L_{i-1})=r_{i-1}\left(p_i^{-1}-1\right)$, $E(r_{i-1}u_i r_{j-1}u_j|L_{i-1})=0$ for $i<j$, we have

$$
\begin{aligned}
V_{LL,t} &= E\left[\operatorname{diag}\left\{\left(p_1^{-1}-1\right)\pi_0^{-1}L_0L_0^{\mathrm{T}},\ldots,\left(p_t^{-1}-1\right)\pi_{t-1}^{-1}L_{t-1}L_{t-1}^{\mathrm{T}}\right\}\right],\\
V_{LS,t} &= E\left[\operatorname{diag}\left\{\left(p_1^{-1}-1\right)p_1L_0h_1^{\mathrm{T}},\ldots,\left(p_t^{-1}-1\right)p_tL_{t-1}h_t^{\mathrm{T}}\right\}\right],\\
V_{SS,t} &= E\left[\operatorname{diag}\left\{\left(p_1^{-1}-1\right)p_1\pi_1h_1h_1^{\mathrm{T}},\ldots,\left(p_t^{-1}-1\right)p_t\pi_th_th_t^{\mathrm{T}}\right\}\right],\\
V_{LY,t} &= \left[E\left\{\left(p_1^{-1}-1\right)\pi_0^{-1}L_0^{\mathrm{T}}r_tY_t/\pi_t\right\},\ldots,E\left\{\left(p_t^{-1}-1\right)\pi_{t-1}^{-1}L_{t-1}^{\mathrm{T}}r_tY_t/\pi_t\right\}\right]^{\mathrm{T}},\\
V_{SY,t} &= \left[E\left\{\left(p_1^{-1}-1\right)p_1h_1^{\mathrm{T}}r_tY_t/\pi_t\right\},\ldots,E\left\{\left(p_t^{-1}-1\right)p_th_t^{\mathrm{T}}r_tY_t/\pi_t\right\}\right]^{\mathrm{T}}.
\end{aligned}
$$

All the V matrices or vectors can be naturally written as formed by diagonal blocks. If V is a matrix, then $V=\operatorname{diag}\{V(1),\ldots,V(t)\}$; if V is a vector, then $V=\{V(1)^{\mathrm{T}},\ldots,V(t)^{\mathrm{T}}\}^{\mathrm{T}}$. Therefore $B_t^*=(B_{1t}^{*\mathrm{T}},\ldots,B_{1t}^{*\mathrm{T}})^{\mathrm{T}}$, where

$$
\begin{aligned}
B_{jt}^* &= \{V_{LL,t}(j)-V_{LS,t}(j)V_{SS,t}^{-1}(j)V_{SL,t}(j)\}^{-1}\{V_{LY,t}(j)-V_{LS,t}(j)V_{SS,t}(j)^{-1}V_{SY,t}(j)\}\\
&= (I_{\dim(L_{j-1})},0)E\left\{\left(p_j^{-1}-1\right)\begin{pmatrix}\pi_{j-1}^{-1}L_{j-1}\\ p_jh_j\end{pmatrix}\begin{pmatrix}L_{j-1}\\ \pi_jh_j\end{pmatrix}^{\mathrm{T}}\right\}^{-1}E\left\{\left(p_j^{-1}-1\right)\begin{pmatrix}\pi_{j-1}^{-1}L_{j-1}\\ p_jh_j\end{pmatrix}\frac{r_tY_t}{\pi_t}\right\}.
\end{aligned}
$$

Proof of Theorem 2

We can write $\hat{Y}_{t,\mathrm{opt}}=\tilde{E}(r_tY_t/\hat{\pi}_t)-\tilde{E}(\hat{\xi}_{t-1})^{\mathrm{T}}\hat{B}_t-\tilde{E}(\hat{\psi}_{t-1})^{\mathrm{T}}\hat{C}_t$. Let $\gamma=(B^{\mathrm{T}},C^{\mathrm{T}},\phi^{\mathrm{T}})^{\mathrm{T}}$ and $\gamma^*=\{B_t^{*\mathrm{T}},C_t^{*\mathrm{T}},(\bar{\phi}_t^*)^{\mathrm{T}}\}^{\mathrm{T}}$, where

$$
\begin{pmatrix}B_t^*\\ C_t^*\end{pmatrix}=\begin{pmatrix}V_{LL,t} & V_{LS,t}\\ V_{SL,t} & V_{SS,t}\end{pmatrix}^{-1}\begin{pmatrix}V_{LY,t}\\ V_{SY,t}\end{pmatrix}. \tag{A1}
$$

Consider $\mu_t(\gamma)=E_\gamma(r_tY_t/\pi_t)-B^{\mathrm{T}}E_\gamma(\xi_{t-1})-C^{\mathrm{T}}E_\gamma(\psi_{t-1})$. Then, under Conditions A1–A4, we have

$$
\{\partial\mu_t(\gamma)/\partial B\}\Big|_{\gamma=\gamma^*}=E_{\gamma^*}(\xi_{t-1})|_{\gamma=\gamma^*}=0,\quad \{\partial\mu_t(\gamma)/\partial C\}\Big|_{\gamma=\gamma^*}=E_{\gamma^*}(\psi_{t-1})|_{\gamma=\gamma^*}=0.
$$

Under Conditions A1–A6, by the results shown in the proof of Theorem 1, we have

$$
\{\partial\mu_t(\gamma)/\partial\phi\}\Big|_{\gamma=\gamma^*}=V_{YS,t}-B_t^{*\mathrm{T}}V_{LS,t}-C_t^{*\mathrm{T}}V_{SS,t}.
$$

To show that $\{\partial\mu_t(\gamma)/\partial\phi\}\Big|_{\gamma=\gamma^*}=0$, it suffices to show that

$$
V_{SS,t}^{-1}V_{SY,t}-V_{SS,t}^{-1}V_{SL,t}B_t^*=C_t^*. \tag{A2}
$$

Matrix algebra for the inverse of a partitioned matrix enables us to establish that

$$
V_{SS,t}^{-1}V_{SY,t}=(V_{SS,t}^{-1}V_{SL,t},I)\begin{pmatrix}V_{LL,t} & V_{LS,t}\\ V_{SL,t} & V_{SS,t}\end{pmatrix}^{-1}\begin{pmatrix}V_{LY,t}\\ V_{SY,t}\end{pmatrix},
$$

which, together with (A1), leads to (A2). Therefore, the Randles (1982) condition is satisfied, and

$$
\hat{Y}_{t,\mathrm{opt}}=\tilde{E}(r_tY_t/\pi_t)-\sum_{j=1}^{t}D_{jt}^{*\mathrm{T}}\tilde{E}\left\{r_{j-1}u_j\begin{pmatrix}\pi_{j-1}^{-1}L_{j-1}\\ p_jh_j\end{pmatrix}\right\}+o_p(n^{-1/2}),
$$

where D_{jt}^* is given in (16). Let η_t be the random quantity as given in (18), then $\hat{Y}_{t,\mathrm{opt}}=\tilde{E}(\eta_t)+o_p(n^{-1/2})$. Because η_t has a second moment, by central limit theorem, (19) holds. Now we shall show that $\operatorname{var}(\eta_t)$ can

be consistently estimated by $\hat{V} = (n-1)^{-1}\tilde{E}\{\hat{\eta}_t - \tilde{E}(\hat{\eta}_t)\}^2$, where

$$\hat{\eta}_{it} = r_{it}Y_{it}/\hat{\pi}_{it} - \sum_{j=1}^{t} \hat{D}_{jt}^{\mathrm{T}} r_{i,j-1}\hat{u}_{ij}(\hat{\pi}_{i,j-1}^{-1}L_{i,j-1}^{\mathrm{T}},\ \hat{p}_{ij}\hat{h}_{ij}^{\mathrm{T}})^{\mathrm{T}}.$$

Since we have already shown that $\tilde{E}(\hat{\eta}_t) = \tilde{E}(\eta_t) + o_p(n^{-1/2})$, it remains to show that $\tilde{E}(\hat{\eta}_t\hat{\eta}_t^{\mathrm{T}}) = \tilde{E}(\eta_t\eta_t^{\mathrm{T}}) + o_p(1)$. By Conditions A1, A2, A4 and A6 there exists a neighbourhood $\bar{\mathcal{N}}_t$ of $\bar{\phi}_t^*$ such that $E(\sup_{\bar{\phi}_t\in\bar{\mathcal{N}}_t}\|\eta_t\|) < \infty$ and $E(\sup_{\bar{\phi}_t\in\bar{\mathcal{N}}_t}\|\eta_t\eta_t^{\mathrm{T}}\|) < \infty$. By Lemma 4.3 of Newey & McFadden (1994), we have $\hat{D}_{jt} = D_{jt}^* + o_p(1)$ and $\tilde{E}(\hat{\eta}_t\hat{\eta}_t^{\mathrm{T}}) = \tilde{E}(\eta_t\eta_t^{\mathrm{T}}) + o_p(1)$. Therefore, $\hat{V} = (n-1)^{-1}\tilde{E}\{\eta_t - \tilde{E}(\eta_t)\}^2 + o_p(n^{-1}) = n^{-1}\mathrm{var}(\eta_t) + o_p(n^{-1})$. That is, $\hat{V}/\{\mathrm{var}(\eta_t)/n\} = 1 + o_p(1)$ and (20) holds.

Sketch of proof for Corollary 1

In light of the arguments in proving Theorem 1, under Conditions A1–A6, 1 and 2, we have

$$\begin{aligned}\tilde{E}_A(\hat{\xi}_{t-1}) &= \tilde{E}_A(\xi_{t-1}) - E\{\tilde{E}_A(\xi_{t-1}\psi_{t-1}^{\mathrm{T}})|\mathcal{F}_N\}E\{\tilde{E}_A(\psi_{t-1}\psi_{t-1}^{\mathrm{T}}/w)\mid\mathcal{F}_N\}^{-1}\\ &\quad\times\tilde{E}_A(\psi_{t-1}/w) + o_p(n^{-1/2}),\\ \tilde{E}_A(r_tY_t/\hat{\pi}_t) &= \tilde{E}_A(r_tY_t/\pi_t) - E[\tilde{E}_A\{(r_tY_t/\pi_t)\psi_{t-1}^{\mathrm{T}}\}\mid\mathcal{F}_N]E\{\tilde{E}_A(\psi_{t-1}\psi_{t-1}^{\mathrm{T}}/w)\mid\mathcal{F}_N\}^{-1}\\ &\quad\times\tilde{E}_A(\psi_{t-1}/w) + o_p(n^{-1/2}).\end{aligned}$$

Since $E\{(r_{it}/p_{it}-1)r_{i,t-1}\mid A,\mathcal{F}_N\} = E[E\{(r_{it}/p_{it}-1)r_{i,t-1}\mid r_{i,t-1},A,\mathcal{F}_N\}\mid A,\mathcal{F}_N] = 0$, we have $E\{\tilde{E}_A(\xi_{t-1})\mid A,\mathcal{F}_N\} = 0$, $E\{\tilde{E}_A(\psi_{t-1}/w)\mid A,\mathcal{F}_N\} = 0$, and

$$\begin{aligned}\mathrm{cov}\{\tilde{E}_A(\xi_{t-1}),\tilde{E}_A(\psi_{t-1}/w)|\mathcal{F}_N\} &= E\{\tilde{E}_A(\xi_{t-1}\psi_{t-1}^{\mathrm{T}})|\mathcal{F}_N\},\\ \mathrm{var}\{\tilde{E}_A(\psi_{t-1})/w|\mathcal{F}_N\} &= E\{\tilde{E}_A(\psi_{t-1}\psi_{t-1}^{\mathrm{T}}/w)|\mathcal{F}_N\},\\ \mathrm{cov}\{\tilde{E}_A(r_tY_t/\pi_t),\tilde{E}_A(\psi_{t-1}/w)|\mathcal{F}_N\} &= E[\tilde{E}_A\{(r_tY_t/\pi_t)\psi_{t-1}^{\mathrm{T}}\}|\mathcal{F}_N].\end{aligned}$$

The rest of this proof follows similarly from the proof of Theorem 1. One important step is to calculate $\mathrm{var}[\tilde{E}_A\{(\xi_{t-1}^{\mathrm{T}},\psi_{t-1}^{\mathrm{T}}/w)^{\mathrm{T}}\}|\mathcal{F}_N]$ and $\mathrm{cov}[\tilde{E}_A(r_tY_t/\pi_t),\tilde{E}_A\{(\xi_{t-1}^{\mathrm{T}},\psi_{t-1}^{\mathrm{T}}/w)^{\mathrm{T}}\}|\mathcal{F}_N]$. Since

$$\begin{aligned}\mathrm{var}\left\{\tilde{E}_A\begin{pmatrix}\xi_{t-1}\\ \psi_{t-1}/w\end{pmatrix}\middle|\mathcal{F}_N\right\} &= \mathrm{var}\left[E\left\{\tilde{E}_A\begin{pmatrix}\xi_{t-1}\\ \psi_{t-1}/w\end{pmatrix}\middle|A,\mathcal{F}_N\right\}\middle|\mathcal{F}_N\right]\\ &\quad + E\left[\mathrm{var}\left\{\tilde{E}_A\begin{pmatrix}\xi_{t-1}\\ \psi_{t-1}/w\end{pmatrix}\middle|A,\mathcal{F}_N\right\}\middle|\mathcal{F}_N\right],\end{aligned}$$

we have to calculate only the second term as the first term equals zero. For the second term,

$$\begin{aligned}\mathrm{var}\left\{\tilde{E}_A\begin{pmatrix}\xi_{t-1}\\ \psi_{t-1}/w\end{pmatrix}\middle|A,\mathcal{F}_N\right\} &= \frac{1}{N^2}\sum_{i\in A}w_i^2\mathrm{var}\left\{\begin{pmatrix}\xi_{i,t-1}\\ \psi_{i,t-1}/w_i\end{pmatrix}\middle|A,\mathcal{F}_N\right\}\\ &= \frac{1}{N^2}\sum_{i\in A}\begin{pmatrix}w_i^2\mathrm{var}(\xi_{i,t-1}|A,\mathcal{F}_N) & w_i\mathrm{cov}(\xi_{i,t-1},\psi_{i,t-1}|A,\mathcal{F}_N)\\ w_i\mathrm{cov}(\psi_{i,t-1},\xi_{i,t-1}|A,\mathcal{F}_N) & \mathrm{var}(\psi_{i,t-1}|A,\mathcal{F}_N)\end{pmatrix}.\end{aligned}$$

Again $\mathrm{var}(\xi_{i,t-1}|A,\mathcal{F}_N)$ can be written as a matrix of diagonal blocks such that it is equal to

$$\mathrm{diag}\{\mathrm{var}(r_{i0}u_{i1}L_{i0}/\pi_{i0}\mid A,\mathcal{F}_N),\ldots,\mathrm{var}(r_{i,t-1}u_{it}L_{i,t-1}/\pi_{i,t-1}\mid A,\mathcal{F}_N)\},$$

where $\mathrm{var}(r_{i,j-1}u_{ij}L_{i,j-1}/\pi_{i,j-1}|A,\mathcal{F}_N) = L_{i,j-1}L_{i,j-1}^{\mathrm{T}}(1/p_{ij}-1)/\pi_{i,j-1}$. Other related terms can be obtained similarly. Thus $\mathrm{cov}[\tilde{E}_A\{(\xi_{t-1}^{\mathrm{T}},\psi_{t-1}^{\mathrm{T}}/w)^{\mathrm{T}}\},\tilde{E}_A(r_tY_t/\pi_t)\mid\mathcal{F}_N] = (\tilde{V}_{LY,t}^{\mathrm{T}},\tilde{V}_{SY,t}^{\mathrm{T}})^{\mathrm{T}}$ and

$$\mathrm{var}\left\{\tilde{E}_A\begin{pmatrix}\xi_{t-1}\\ \psi_{t-1}/w\end{pmatrix}\middle|\mathcal{F}_N\right\} = \begin{pmatrix}\tilde{V}_{LL,t} & \tilde{V}_{LS,t}\\ \tilde{V}_{SL,t} & \tilde{V}_{SS,t}\end{pmatrix},$$

where

$$\tilde{V}_{LL,t} = N^{-2} \sum_{i=1}^{N} w_i \mathrm{diag}\left\{(p_{i1}^{-1} - 1)\pi_{i0}^{-1} L_{i0} L_{i0}^{\mathrm{T}}, \ldots, (p_{it}^{-1} - 1)\pi_{i,t-1}^{-1} L_{i,t-1} L_{i,t-1}^{\mathrm{T}}\right\},$$

$$\tilde{V}_{LS,t} = N^{-2} \sum_{i=1}^{N} \mathrm{diag}\left\{(p_{i1}^{-1} - 1) p_{i1} L_{i0} h_{i1}^{\mathrm{T}}, \ldots, (p_{it}^{-1} - 1) p_{it} L_{i,t-1} h_{it}^{\mathrm{T}}\right\},$$

$$\tilde{V}_{SS,t} = N^{-2} \sum_{i=1}^{N} w_i^{-1} \mathrm{diag}\left\{(p_{i1}^{-1} - 1) p_{i1} \pi_{i,1} h_{i1} h_{i1}^{\mathrm{T}}, \ldots, (p_{it}^{-1} - 1) p_{it} \pi_{it} h_{it} h_{it}^{\mathrm{T}}\right\}$$

$$\tilde{V}_{LY,t} = N^{-2} \sum_{i=1}^{N} w_i Y_{it} \left\{(p_{i1}^{-1} - 1)\pi_{i0}^{-1} L_{i0}^{\mathrm{T}}, \ldots, (p_{it}^{-1} - 1)\pi_{i,t-1}^{-1} L_{i,t-1}^{\mathrm{T}}\right\}^{\mathrm{T}}$$

$$\tilde{V}_{SY,t} = N^{-2} \sum_{i=1}^{N} Y_{it} \left\{(1 - p_{i1}) h_{i1}^{\mathrm{T}}, \ldots, (1 - p_{it}) h_{it}^{\mathrm{T}}\right\}^{\mathrm{T}}.$$

In a way similar to the diagonal blockwise technique used in the proof of Theorem 1, we obtain the optimal $B_t^* = (B_{1t}^{*\mathrm{T}}, \ldots, B_{tt}^{*\mathrm{T}})^{\mathrm{T}}$, where $B_{jt}^* = (I_{\dim(L_{j-1})}, 0) D_{jt}^*$ and

$$D_{jt}^* = \tilde{E}_{\mathcal{F}}\left\{w\left(p_j^{-1} - 1\right)\begin{pmatrix}\pi_{j-1}^{-1} L_{j-1} \\ p_j h_j / w\end{pmatrix}\begin{pmatrix}L_{j-1} \\ \pi_j h_j / w\end{pmatrix}^{\mathrm{T}}\right\}^{-1} \tilde{E}_{\mathcal{F}}\left\{w\left(p_j{}^{-1} - 1\right)\begin{pmatrix}\pi_{j-1}^{-1} L_{j-1} \\ p_j h_j / w\end{pmatrix} Y_t\right\}.$$

Therefore, the consistency of the estimator in (24) follows.

References

Baker, W. K. (1995). Allen and Meyer's 1990 longitudinal study: A reanalysis and reinterpretation using structural equation modeling. *Hum. Relations* **48**, 169–86.

Beaumont, J.-F. & Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Can. J. Statist.* **37**, 400–16.

Bollinger, C. R. & David, M. H. (1997). Modeling discrete choice with response error: Food stamp participation. *J. Am. Statist. Assoc.* **92**, 827–35.

Bollinger, C. R. & David, M. H. (2001). Estimation with response error and nonresponse: Food-stamp participation in the SIPP. *J. Bus. Econ. Statist.* **19**, 129–41.

Cao, W., Tsiatis, A. & Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96**, 723–4.

Diggle, P. J., Heagerty, P., Liang, K.-Y. & Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford: Oxford University Press, 2nd ed.

Diggle, P. J. & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *Appl. Statist.* **43**, 49–94.

Duncan, K. B. & Stasny, E. A. (2001). Using propensity scores to control coverage bias in telephone surveys. *Survey Methodol.* **27**, 121–30.

Ekholm, A. & Laaksonen, S. (1991). Weighting via response modeling in the Finnish Household Budget Survey. *J. Offic. Statist.* **7**, 325–7.

Fay, R. E. (1992). When are inferences from multiple imputation valid? In *Proc. Survey Res. Meth. Sect., Am. Statist. Assoc.* Washington DC: American Statistical Association, pp. 227–32.

Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G. (2008). *Longitudinal Data Analysis*. Boca Raton: Chapman & Hall/CRC.

Flanders, W. D. & Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Statist. Med.* **10**, 739–47.

Fuller, W. A., Loughin, M. M. & Baker, H. D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 national food consumption survey. *Survey Methodol.* **20**, 75–85.

Henmi, M. & Eguchi, S. (2004). A paradox concerning nuisance parameters and projected estimating functions. *Biometrika* **91**, 929–41.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc.* **47**, 663–85.

IBRAHIM, J. G., CHEN, M.-H. & LIPSITZ, S. R. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika* **88**, 551–64.
KALTON, G. & KASPRZYK, D. (1986). The treatment of missing data. *Survey Methodol.* **12**, 1–16.
KIM, J. K. & KIM, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Can. J. Statist.* **35**, 501–14.
KIM, J. K., NAVARRO, A. & FULLER, W. A. (2006). Replication variance estimation for two-phase stratified sampling. *J. Am. Statist. Assoc.* **101**, 312–20.
KIM, J. K. & RAO, J. N. K. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika* **96**, 917–32.
KORINEK, A., MISTIAEN, J. A. & RAVALLION, M. (2007). An econometric method of correcting for unit nonresponse bias in surveys. *J. Economet.* **136**, 213–35.
LAAKSONEN, S. (2007). Weighting for two-phase surveyed data. *Survey Methodol.* **33**, 121–30.
LAAKSONEN, S. & CHAMBERS, R. L. (2006). Survey estimation under informative nonresponse with follow-up. *J. Offic. Statist.* **22**, 81–95.
LITTLE, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *J. Am. Statist. Assoc.* **90**, 1112–21.
LITTLE, R. J. A. & VARTIVARIAN, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodol.* **31**, 161–8.
MANSKI, C. F. & LERMAN, S. R. (1977). The estimation of choice probabilities from choice based samples. *Econometrica* **45**, 1977–88.
MOLENBERGHS, G. & KENWARD, M. G. (2007). *Missing Data in Clinical Studies*. New York: Wiley.
MOLENBERGHS, G., KENWARD, M. G. & LESAFFRE, E. (1997). The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika* **84**, 33–44.
NEWEY, W. K. & MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, vol. 4, ch. 36, 1st ed., pp. 2111–245. Amsterdam: Elsevier.
PIERCE, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *Ann. Statist.* **10**, 475–8.
RANDLES, R. H. (1982). On the asymptotic normality of statistics with estimated parameters. *Ann. Statist.* **10**, 462–74.
RIZZO, L., KALTON, G. & BRICK, J. M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodol.* **22**, 43–53.
ROBINS, J. M., ROTNITZKY, A. & ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846–66.
ROBINS, J. M., ROTNITZKY, A. & ZHAO, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Statist. Assoc.* **90**, 106–21.
ROSENBAUM, P. R. (1987). Model-based direct adjustment. *J. Am. Statist. Assoc.* **82**, 387–94.
ROSENBAUM, P. R. & RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.
ROTNITZKY, A., ROBINS, J. M. & SCHARFSTEIN, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J. Am. Statist. Assoc.* **93**, 1321–39.
RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–90.
SHAO, J. & STEEL, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *J. Am. Statist. Assoc.* **94**, 254–65.
SLUD, E. V. & BAILEY, L. (2010). Evaluation and selection of models for attrition nonresponse adjustment. *J. Offic. Statist.* **26**, 1–18.
TAN, Z. (2006). A distributional approach for causal inference using propensity scores. *J. Am. Statist. Assoc.* **101**, 1619–37.

[*Received May* 2011. *Revised March* 2012]