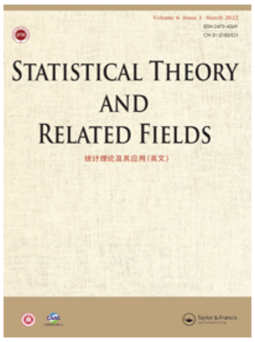

Taylor & Francis
Taylor & Francis Group

Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Nonignorable item nonresponse in panel data

Sijing Li & Jun Shao

To cite this article: Sijing Li & Jun Shao (2022) Nonignorable item nonresponse in panel data, Statistical Theory and Related Fields, 6:1, 58-71, DOI: 10.1080/24754269.2020.1856591

To link to this article: https://doi.org/10.1080/24754269.2020.1856591

Published online: 17 Dec 2020.

STATISTICAL THEORY AND RELATED FIELDS
2022, VOL. 6, NO. 1, 58–71
https://doi.org/10.1080/24754269.2020.1856591

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Nonignorable item nonresponse in panel data

Sijing Li and Jun Shao

Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA

ABSTRACT

To estimate unknown population parameters based on panel data having nonignorable item nonresponse, we propose an innovative data grouping approach according to the number of observed components in the multivariate outcome $\boldsymbol{y}$ when the joint distribution of $\boldsymbol{y}$ and associated covariate $\boldsymbol{x}$ is nonparametric and the nonresponse probability conditional on $\boldsymbol{y}$ and $\boldsymbol{x}$ has a parametric form. To deal with the identifiability issue, we utilise a nonresponse instrument $\boldsymbol{z}$, an auxiliary variable related to $\boldsymbol{y}$ but not related to the nonresponse probability conditional on $\boldsymbol{y}$ and $\boldsymbol{x}$. We apply a modified generalised method of moments to obtain estimators of the parameters in the nonresponse probability, and a generalised regression estimation to utilise covariate information for efficient estimation of population parameters. Consistency and asymptotic normality of the proposed estimators of the population parameters are established. Simulation and real data results are presented.

ARTICLE HISTORY
Received 11 June 2020
Revised 16 August 2020
Accepted 24 November 2020

KEYWORDS
Cluster sampling; generalised regression; grouping; homogeneity; multivariate outcome

1. Introduction

Panel data are collected in many statistical applications, such as sample surveys, clinical trials, economics and social sciences. For example, cluster sampling results in panel data, which occurs in social studies and sample surveys when mutual homogeneity within clusters is evident in the population of interest. Multivariate outcome from a single sampled unit also leads to panel data.

Item nonresponse is a common phenomena in panel data, i.e., some components of the panel, not necessary the entire panel, may be missing. For example, in survey studies, subjects may not respond to all questions; in cluster sampling, some units within a cluster may not respond; in multivariate outcome, some components are measured while the others are not. Estimation and statistical inference without taking nonresponse into consideration could lead to seriously biased estimators and conclusions.

Consider a k-dimensional response or outcome vector $\boldsymbol{y}$ of interest that is subject to item nonresponse. Let $\boldsymbol{r}$ be the response indicator vector of $\boldsymbol{y}$, i.e., the jth component of $\boldsymbol{r}$ is 1 (or 0) if the jth component of $\boldsymbol{y}$ is observed (or not observed), $j = 1, \ldots, k$. Statistical approaches dealing with missing data usually depend on the nonresponse propensity (or mechanism), i.e., the conditional distribution of $\boldsymbol{r}$ given $(\boldsymbol{y}, \boldsymbol{x})$, denoted by $p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{x})$, where $\boldsymbol{x}$ is a covariate vector associated with $\boldsymbol{y}$ and is always observed. If $p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{x}) = p(\boldsymbol{r} \mid \boldsymbol{y}_o, \boldsymbol{x})$, where $\boldsymbol{y}_o$ is the observed part of $\boldsymbol{y}$, then nonresponse is ignorable (Little & Rubin, 2002; Rubin, 1976). Otherwise, nonresponse is nonignorable. While there is a rich literature for valid inference on unknown $p(\boldsymbol{y})$ (the distribution of $\boldsymbol{y}$) or $p(\boldsymbol{y} \mid \boldsymbol{x})$ (the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$) under ignorable nonresponse (S. X. Chen et al., 2008; Little & Rubin, 2002; Robins & Rotiv, 1997; Rotnitzky & Robins, 1997; Rubin, 1976), statistical inference faces serious challenges under nonignorable nonresponse when $p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{x})$ depends on $\boldsymbol{y}$ as well as some components of $\boldsymbol{x}$.

We provide a brief review of the progress in research on general nonignorable nonresponse in $\boldsymbol{y}$. Greenlees et al. (1982) proposed to handle nonignorable item nonresponse by maximum likelihood estimation, assuming both $p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{x})$ and $p(\boldsymbol{y} \mid \boldsymbol{x})$ are parametric; however, the non-identifiability issue caused by nonignorable nonresponse is not well-addressed and, thus, the result is not rigorous. Besides, a fully parametric approach is sensitive to the parametric model assumptions. Since the population is not identifiable when both $p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{x})$ and $p(\boldsymbol{y} \mid \boldsymbol{x})$ are nonparametric (Robins & Rotiv, 1997), efforts have been made in situations where one of $p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{x})$ and $p(\boldsymbol{y} \mid \boldsymbol{x})$ is parametric or semiparametric. Tang et al. (2003) considered the situation where $p(\boldsymbol{y} \mid \boldsymbol{x})$ is parametric but $p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{x})$ is nonparametric, and provided a rigorous treatment of the identifiability issue for the first time; but they assumed that the nonresponse propensity depends only on $\boldsymbol{y}$, i.e., $p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{x}) = p(\boldsymbol{r} \mid \boldsymbol{y})$, which may be impractical. This result was extended by Zhao and Shao (2015) to more realistic situation where $p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{x}) = p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{u})$ and $\boldsymbol{u}$ is a sub-vector of $\boldsymbol{x}$. While both previously cited

CONTACT Sijing Li lsabrina126@gmail.com, sli394@wisc.edu

papers assumed a parametric $p(\boldsymbol{y}\,|\,\boldsymbol{x})$ but a unspecified $p(\boldsymbol{r}\,|\,\boldsymbol{y},\boldsymbol{x})$, parallel results were established by Wang et al. (2014) and J. Shao and L. Wang (2016) under a univariate $\boldsymbol{y}$ ($k=1$) with a nonparametric $p(\boldsymbol{y}\,|\,\boldsymbol{x})$ and a parametric or semi-parametric $p(\boldsymbol{r}\,|\,\boldsymbol{y},\boldsymbol{x})$, which are particularly useful in sample surveys where it is difficult to find a suitable parametric model for $p(\boldsymbol{y}\,|\,\boldsymbol{x})$. Other than the results under a parametric model on $p(\boldsymbol{y}\,|\,\boldsymbol{x})$, there is no general result on multivariate $\boldsymbol{y}$ having nonignorable item nonresponse, although Wu and Carroll (1988), Xu and Shao (2009), and Shao and Zhang (2015) obtained some results when the dependence of $\boldsymbol{r}$ on $\boldsymbol{y}$ is through an unobserved random effect $\boldsymbol{b}$, i.e., $p(\boldsymbol{r}\,|\,\boldsymbol{y},\boldsymbol{x})=p(\boldsymbol{r}\,|\,\boldsymbol{b},\boldsymbol{x})$.

Under nonparametric $p(\boldsymbol{y})$ and $p(\boldsymbol{y}\,|\,\boldsymbol{x})$, in this paper we propose an innovative data grouping approach to construct valid estimators of population parameters in the presence of nonignorable item nonresponse in $\boldsymbol{y}$, assuming the following two main assumptions.

(A1) Given $(\boldsymbol{y},\boldsymbol{x})$, components of $\boldsymbol{r}$ are conditionally independent and identically distributed.

(A2) Given $(\boldsymbol{y},\boldsymbol{x})$, the conditional probability of observing a component of $\boldsymbol{y}$ is $\pi_\theta(\boldsymbol{y},\boldsymbol{u})$, where π_θ is a parametric function of $(\boldsymbol{y},\boldsymbol{u})$ with an unknown parameter vector θ and $\boldsymbol{x}=(\boldsymbol{u},\boldsymbol{z})$ with $p(\boldsymbol{y}\,|\,\boldsymbol{x})$ depending on $\boldsymbol{z}$.

Our main methodology is introduced in Section 2, followed by some simulation results in Section 3 and two real data examples in Section 4. Section 5 contains some technical proofs.

2. Methodology

We use the notation developed in Section 1. Our inference is based on a training sample of size n, $(\boldsymbol{y}_i,\boldsymbol{x}_i,\boldsymbol{r}_i)$, $i=1,\ldots,n$, which are independent and identically distributed with $(\boldsymbol{y},\boldsymbol{x},\boldsymbol{r})$. Values of $\boldsymbol{x}_i$ are always observed and components of $\boldsymbol{y}_i$ are observed if and only if the corresponding components of $\boldsymbol{r}_i$ are equal to 1.

2.1. Grouping

When there is no nonresponse, values in the entire set $\{(\boldsymbol{y}_i,\boldsymbol{x}_i),\ i=1,\ldots,n\}$ are exchangeable. But this does not hold in the presence of nonignorable item nonresponse in $\boldsymbol{y}$. Although $(\boldsymbol{y}_i,\boldsymbol{x}_i)$'s with the same nonresponse pattern $\boldsymbol{r}_i$ are exchangeable, there are a total of 2^k different nonresponse patterns when k is the dimension of $\boldsymbol{y}$. Thus although grouping according to nonresponse patter is natural to achieve within-group homogeneity, each group may not have enough units for efficient estimation or inference.

Our main idea is to divide data into $k+1$ groups with within-group homogeneity, using the following key lemma under assumption (A1).

Lemma 2.1. *Let Δ be the number of observed components in $\boldsymbol{y}$. Under (A1), $p(\boldsymbol{y},\boldsymbol{x}\,|\,\boldsymbol{r})=p(\boldsymbol{y},\boldsymbol{x}\,|\,\Delta)$, i.e., the conditional distribution of $(\boldsymbol{y},\boldsymbol{x})$ given $\boldsymbol{r}$ is the same as the conditional distribution of $(\boldsymbol{y},\boldsymbol{x})$ given Δ.*

Proof: Let $\pi=\pi(\boldsymbol{y},\boldsymbol{x})$ be the conditional probability of observing a component of $\boldsymbol{y}$ given $(\boldsymbol{y},\boldsymbol{x})$. Under (A1), $p(\boldsymbol{r}\,|\,\boldsymbol{y},\boldsymbol{x})=\pi^\Delta(1-\pi)^{k-\Delta}$ and Δ follows a binomial distribution with probability π and size k conditioned on $(\boldsymbol{y},\boldsymbol{x})$. The result follows from

$$
\begin{aligned}
&\frac{p(\boldsymbol{y},\boldsymbol{x}\,|\,\boldsymbol{r})}{p(\boldsymbol{y},\boldsymbol{x}\,|\,\Delta)}\\
&=\frac{p(\boldsymbol{r}\,|\,\boldsymbol{y},\boldsymbol{x})p(\boldsymbol{y},\boldsymbol{x})}{\int\int p(\boldsymbol{r}\,|\,\boldsymbol{y},\boldsymbol{x})p(\boldsymbol{y},\boldsymbol{x})\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{x}}\\
&\quad\Big/\frac{p(\Delta\,|\,\boldsymbol{y},\boldsymbol{x})p(\boldsymbol{y},\boldsymbol{x})}{\int\int p(\Delta\,|\,\boldsymbol{y},\boldsymbol{x})p(\boldsymbol{y},\boldsymbol{x})\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{x}}\\
&=\frac{p(\boldsymbol{r}\,|\,\boldsymbol{y},\boldsymbol{x})}{p(\Delta\,|\,\boldsymbol{y},\boldsymbol{x})}\cdot\frac{\int\int p(\Delta\,|\,\boldsymbol{y},\boldsymbol{x})p(\boldsymbol{y},\boldsymbol{x})\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{x}}{\int\int p(\boldsymbol{r}\,|\,\boldsymbol{y},\boldsymbol{x})p(\boldsymbol{y},\boldsymbol{x})\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{x}}\\
&=\frac{\pi^\Delta(1-\pi)^{k-\Delta}}{\binom{k}{\Delta}\pi^\Delta(1-\pi)^{k-\Delta}}\\
&\qquad\times\frac{\binom{k}{\Delta}\int\int\pi^\Delta(1-\pi)^{k-\Delta}p(\boldsymbol{y},\boldsymbol{x})\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{x}}{\int\int\pi^\Delta(1-\pi)^{k-\Delta}p(\boldsymbol{y},\boldsymbol{x})\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{x}}\\
&=1.
\end{aligned}
$$

According to Lemma 2.1, we can partition the whole dataset into $k+1$ groups, $\{(\boldsymbol{y}_i,\boldsymbol{x}_i),\Delta_i=d\}$, $d=0,1,\ldots,k$, where Δ_i is the number of observed components in $\boldsymbol{y}_i$. Each group $\{(\boldsymbol{y}_i,\boldsymbol{x}_i),\Delta_i=d\}$ contains exchangeable values and enough units for inference as long as k is much smaller than n. ∎

2.2. Estimation under cluster sampling

We consider the situation where components of $\boldsymbol{y}$ have the same distribution (e.g., we have panel data under cluster sampling) and estimation of a parameter in the population of $\boldsymbol{y}$ is our interest. To illustrate, we focus on the estimation of μ, the mean of a component of $\boldsymbol{y}$. For $d=1,\ldots,k$, and each group with $\Delta_i=d$, the within-group sample mean of observed values is

$$\bar{y}_d=\frac{1}{dn_d}\sum_{i:\Delta_i=d}\sum_{j=1}^{k}r_{ij}y_{ij},\tag{1}$$

where y_{ij} and r_{ij} are the jth components of $\boldsymbol{y}_i$ and $\boldsymbol{r}_i$, respectively, and n_d is the number of units with $\Delta_i=d$. Each $\bar{y}_d$ is an estimator of $\mu_d=E(y_{ij}\,|\,\Delta_i=d)$. Note that $\bar{y}_0$ is not defined.

If $\mu_0=E(y_{ij}\,|\,\Delta_i=0)$ is known, then the overall population mean

$$\mu=\sum_{d=0}^{k}p_d\mu_d,$$

where $p_d = P(\Delta = d)$, can be estimated by

$$\widetilde{\mu} = \frac{n_0}{n}\mu_0 + \sum_{d=1}^{k} \frac{n_d}{n}\bar{y}_d. \tag{2}$$

The proof of the following result is deferred to Section 5.

Theorem 2.1. *Assume* (A1) *holds and that components of* $\boldsymbol{y}$ *have the same distribution with finite second-order moment Then, as* $n \to \infty$,

$$\sqrt{n}(\widetilde{\mu} - \mu) \to N\left(0, p_0\mu_0^2 + \sum_{d=1}^{k} p_d(\sigma_d^2 + \mu_d^2) - \mu^2\right)$$

in distribution, (3)

where $\sigma_d^2 = \mathrm{Var}(y_{ij} \mid \Delta_i = d)$, $d = 1, \ldots, k$.

Since $\mu_0 = E(y_{ij} \mid \Delta_i = 0)$ is usually unknown, however, $\widetilde{\mu}$ is not an estimator and we need to find a way to estimate μ_0. In the group with $\Delta_i = 0$, all components of $\boldsymbol{y}_i$ are missing. Thus some assumption is needed to relate this group with other groups. Under assumption (A2), our idea is to solve this problem using data in the group with $\Delta_i = k$, the group with completely observed $\boldsymbol{y}_i$'s. From

$$p(\boldsymbol{y},\boldsymbol{x}) = \frac{p(\boldsymbol{y},\boldsymbol{x} \mid \Delta = 0)P(\Delta = 0)}{P(\Delta = 0 \mid \boldsymbol{y},\boldsymbol{x})} = \frac{p(\boldsymbol{y},\boldsymbol{x} \mid \Delta = k)P(\Delta = k)}{P(\Delta = k \mid \boldsymbol{y},\boldsymbol{x})},$$

we obtain the following relationship:

$$\begin{aligned} p(\boldsymbol{y},\boldsymbol{x} \mid \Delta = 0) &= \frac{P(\Delta = k)}{P(\Delta = 0)} \cdot \frac{P(\Delta = 0 \mid \boldsymbol{y},\boldsymbol{x})}{P(\Delta = k \mid \boldsymbol{y},\boldsymbol{x})} p(\boldsymbol{y},\boldsymbol{x} \mid \Delta = k) \\ &= \frac{P(\Delta = k)}{P(\Delta = 0)} \cdot \frac{\{1 - \pi_\theta(\boldsymbol{y},\boldsymbol{u})\}^k}{\{\pi_\theta(\boldsymbol{y},\boldsymbol{u})\}^k} p(\boldsymbol{y},\boldsymbol{x} \mid \Delta = k), \end{aligned}$$

where the second equality follows from (A1)–(A2) and $\pi_\theta(\boldsymbol{y},\boldsymbol{u})$ is defined in (A2), the conditional probability of observing a component of $\boldsymbol{y}$ given $(\boldsymbol{y},\boldsymbol{x})$. The ratio $P(\Delta = k)/P(\Delta = 0)$ can be estimated by n_k/n_0. If we can obtain an estimator $\widehat{\theta}$ of θ, then characteristics in $p(\boldsymbol{y},\boldsymbol{x} \mid \Delta = 0)$ can be estimated using this relationship, n_k/n_0, $\widehat{\theta}$, and estimators of characteristics in $p(\boldsymbol{y},\boldsymbol{x} \mid \Delta = k)$ with completely observed $(\boldsymbol{y},\boldsymbol{x})$.

Thus, $\mu_0 = E(y_{ij} \mid \Delta_i = 0)$ can be estimated by

$$\begin{aligned} \widehat{\mu}_0 &= \frac{n_k}{n_0}\int \frac{\{1 - \pi_{\widehat{\theta}}(\boldsymbol{y},\boldsymbol{u})\}^k}{\{\pi_{\widehat{\theta}}(\boldsymbol{y},\boldsymbol{u})\}^k} y \, \mathrm{d}\widehat{F}_k(\boldsymbol{y},\boldsymbol{x}) \\ &= \frac{1}{kn_0}\sum_{i:\Delta_i = k}\sum_{j=1}^{k} \frac{\{1 - \pi_{\widehat{\theta}}(\boldsymbol{y}_i,\boldsymbol{u}_i)\}^k}{\{\pi_{\widehat{\theta}}(\boldsymbol{y}_i,\boldsymbol{u}_i)\}^k} r_{ij}y_{ij}, \end{aligned}$$

where y is a component of $\boldsymbol{y}$ and $\widehat{F}_k$ is the empirical distribution based on the data set $\{(\boldsymbol{y}_i,\boldsymbol{x}_i), \Delta_i = k\}$.

Once μ_0 is estimated by $\widehat{\mu}_0$, the overall population mean μ can be estimated by

$$\widehat{\mu} = \frac{n_0}{n}\widehat{\mu}_0 + \sum_{d=1}^{k} \frac{n_d}{n}\bar{y}_d. \tag{4}$$

In this way, other population characteristics can be similarly estimated. For example, if we want to estimate the distribution of a component of $\boldsymbol{y}$ at a point t, then we just need to replace y_{ij} by the indicator of $y_{ij} \le t$ in the previous discussion. Quantiles of F can then be estimated. Estimators of correlation between two components of $\boldsymbol{y}$ and between $\boldsymbol{y}$ and $\boldsymbol{x}$ can be similarly derived. We can also estimate parameters via estimating equations.

2.3. Estimation of θ in propensity

To complete our proposed methodology, we need to construct an estimator $\widehat{\theta}$ of θ under (A1)–(A2). To estimate θ, we follow the approach of generalised method of moments (GMM) in Wang et al. (2014) for the univariate response, but add a novel modification by utilising the multivariate structure of $\boldsymbol{y}$.

Define an L-dimensional estimating function

$$G(\theta) = \left(g_1(\boldsymbol{y},\boldsymbol{x},\boldsymbol{r},\theta), \ldots, g_L(\boldsymbol{y},\boldsymbol{x},\boldsymbol{r},\theta)\right)',$$

where a' is the transpose of a, L is an integer $\ge$ the dimension of θ and the form of g_l is specified later. These functions are chosen so that, at the true parameter value θ, $E\{G(\theta)\} = 0$ and $E\{\partial G(\theta)/\partial\theta\}$ is of full rank. Let

$$G_n(\theta) = \left(\frac{1}{n}\sum_{i=1}^{n} g_1(\boldsymbol{y}_i,\boldsymbol{x}_i,\boldsymbol{r}_i,\theta), \ldots, \frac{1}{n}\sum_{i=1}^{n} g_L(\boldsymbol{y}_i,\boldsymbol{x}_i,\boldsymbol{r}_i,\theta)\right)'.$$

If L is the same as the dimension of θ, then we estimate θ by $\widehat{\theta}$ such that $G_n(\widehat{\theta}) = 0$. If L is larger than the dimension of θ, we apply the two-step GMM (Hall, 2005; Hansen, 1982) as follows:

(1) Obtain $\widehat{\theta}^{(1)}$ by minimising $\{G_n(\theta)\}'G_n(\theta)$.
(2) Obtain $\widehat{\theta}$ by minimising $\{G_n(\theta)\}'\widehat{W}G_n(\theta)$, where $\widehat{W}$ is the inverse of $L \times L$ matrix whose (l, m) element is $n^{-1}\sum_{i=1}^{n} g_l(\boldsymbol{y}_i,\boldsymbol{x}_i,\boldsymbol{r}_i,\widehat{\theta}^{(1)})g_m(\boldsymbol{y}_i,\boldsymbol{x}_i,\boldsymbol{r}_i,\widehat{\theta}^{(1)})$.

The optimisation can be solved by using the MATLAB function `fminsearch`, which is applied in our simulation and data analysis in Sections 3 and 4.

It remains to specify the form of $G(\theta)$. Suppose first that $\boldsymbol{z}$ is discrete and has q categories, say $\boldsymbol{z} = 1, \ldots, q$. A straightforward extension of the approach in Wang

Table 1. Example of D_1, D_2, D_3 when $k = 3$ and $n = 30$.

Entire data set				D_1				D_2				D_3			
Unit	y_1	y_2	y_3	Unit	y_1	y_2	y_3	Unit	y_1	y_2	y_3	Unit	y_1	y_2	y_3
1	?	?	?	2	✓	✓	✓	2	✓	✓	✓	2	✓	✓	✓
2	✓	✓	✓	3	?	✓	✓	5	✓	✓	✓	5	✓	✓	✓
3	?	✓	✓	5	✓	✓	✓	8	✓	✓	✓	7	✓	✓	?
4	✓	?	?	8	✓	✓	✓	11	✓	?	✓	8	✓	✓	✓
5	✓	✓	✓	12	✓	✓	✓	12	✓	✓	✓	10	✓	✓	?
6	?	?	✓	15	✓	✓	✓	14	✓	?	✓	12	✓	✓	✓
7	✓	✓	?	16	?	✓	✓	15	✓	✓	✓	15	✓	✓	✓
8	✓	✓	✓	17	✓	✓	✓	17	✓	✓	✓	17	✓	✓	✓
9	?	?	✓	21	✓	✓	✓	21	✓	✓	✓	18	✓	✓	?
10	✓	✓	?	23	✓	✓	✓	22	✓	?	✓	20	✓	✓	?
11	✓	?	✓	24	?	✓	✓	23	✓	✓	✓	21	✓	✓	✓
12	✓	✓	✓	28	✓	✓	✓	27	✓	?	✓	23	✓	✓	✓
13	?	✓	?					28	✓	✓	✓	28	✓	✓	✓
14	✓	?	✓					30	✓	?	✓	29	✓	✓	?
15	✓	✓	✓												
16	?	✓	✓												
17	✓	✓	✓												
18	✓	✓	?												
19	?	?	?												
20	✓	✓	?												
21	✓	✓	✓												
22	✓	?	✓												
23	✓	✓	✓												
24	?	✓	✓												
25	?	?	?												
26	?	✓	?												
27	✓	?	✓												
28	✓	✓	✓												
29	✓	✓	?												
30	✓	?	✓												

et al. (2014) (from univariate response to multivariate $\boldsymbol{y}$) is using

$$G(\theta) = \boldsymbol{v}\left\{\frac{r_1\cdots r_k}{[\pi_\theta(\boldsymbol{y},\boldsymbol{u})]^k} - 1\right\}, \tag{5}$$

where r_j is the jth component of the vector $\boldsymbol{r}$ of response indicators and $\boldsymbol{v}$ is a vector whose first q components are indicators of $\boldsymbol{z} = 1, \ldots, q$ and the last p components are the p-dimensional covariate vector $\boldsymbol{u}$ in (A2). With this choice of G, $E\{G(\theta)\} = 0$ under (A1)–(A2).

However, there are two problems. First, the partially observed responses in $\boldsymbol{y}$ are not used in (5), since $r_1\cdots r_k = 1$ if and only if all components of $\boldsymbol{y}$ are observed. Second, a more serious issue is that L may be smaller than the dimension of θ. For example, if $\boldsymbol{u}$ is continuous and

$$\pi_\theta(\boldsymbol{y},\boldsymbol{u}) = \{1 + \exp(\alpha + \beta'\boldsymbol{y} + \gamma'\boldsymbol{u})\}^{-1}, \tag{6}$$

where $\theta = (\alpha, \beta', \gamma')'$, α is univariate, β is k-dimensional, and γ is p-dimensional, then the dimension of θ is $p+k+1$ and $L = p+q$; in this case $L \geq p+k+1$ requires that $q > k$. That is, we are not able to apply GMM if z does not have more than k categories.

To overcome this difficulty, we consider the following modification. First, we construct k overlapped subsets $D_1, \ldots, D_k$ of the entire data set, where D_h contains data from units whose y_{ih} may be missing but all other components are observed, $h = 1, \ldots, k$. With the notation $r_j =$ the jth component of $\boldsymbol{r}$, $D_h = \{r_1 = \cdots = r_{h-1} = r_{h+1} = \cdots = r_k = 1\}$. Table 1 provides an example of D_1, D_2, D_3, where a check mark indicates an observed datum and a question mark indicates a nonresponse.

For each fixed h, we consider

$$G^{(h)}(\theta) = \boldsymbol{v}_h\left\{\frac{r_h}{\pi_\theta(\boldsymbol{y},\boldsymbol{u})} - 1\right\}, \tag{7}$$

where $L = p+q+k-1$, $\boldsymbol{v}_h$ is the L-dimensional vector whose first $p+q$ components are the same as those of $\boldsymbol{v}$ in (5), the rest $k-1$ components are $r_1y_1, \ldots, r_{h-1}y_{h-1}$, $r_{h+1}y_{h+1}, \ldots, r_ky_k$, and y_h is the hth component of $\boldsymbol{y}$. To see why the function $G^{(h)}(\theta)$ in (7) can be used in estimation equation, note that

$$\begin{aligned} &E\{G^{(h)}(\theta)\} \\ &= E\left\{E\left\{\boldsymbol{v}_h\left[\frac{r_h}{\pi_\theta(\boldsymbol{y},\boldsymbol{u})} - 1\right]\middle|\, \boldsymbol{y},\boldsymbol{u},D_h\right\}\right\} \\ &= E\left\{E(\boldsymbol{v}_h \mid \boldsymbol{y},\boldsymbol{u},D_h)\left[\frac{E(r_h \mid \boldsymbol{y},\boldsymbol{u},D_h)}{\pi_\theta(\boldsymbol{y},\boldsymbol{u})} - 1\right]\right\} \\ &= 0, \end{aligned}$$

where the second equality follows from the independence between z and r_h conditioned on $(\boldsymbol{y},\boldsymbol{u},D_h)$ and the last equality follows from the fact that $E(r_h \mid \boldsymbol{y},\boldsymbol{u},D_h) = E(r_h \mid \boldsymbol{y},\boldsymbol{u}) = \pi_\theta(\boldsymbol{y},\boldsymbol{u})$ under (A1)–(A2).

Note that the key difference between $G(\theta)$ in (5) and $G^{(h)}(\theta)$ in (7) is that the components of $\boldsymbol{y}$ other than the

hth component are used as 'covariates' and included in the vector $\boldsymbol{v}_h$ in (7). In this way, we have more estimating functions and does not need to have the restriction $q > k$ in the case of (6), because $L = p + q + k - 1 \geq p + k + 1$ is easily satisfied as long as $q \geq 2$.

If we apply the GMM algorithm with $G(\theta)$ in (5) replaced by $G^{(h)}(\theta)$ in (7), we can obtain a GMM estimator $\widehat{\theta}^{(h)}$ for every h. Our proposed final GMM estimator of θ is then the weighted average estimator

$$\widehat{\theta} = \sum_{h=1}^{k} m_h \widehat{\theta}^{(h)} \Bigg/ \sum_{h=1}^{k} m_h,$$

where m_h is the number of units in D_h.

When $\boldsymbol{z}$ has continuous components, we can apply the method by discretising $\boldsymbol{z}$ into q categories or use moments of $\boldsymbol{z}$ as components of $\boldsymbol{v}$.

Under the same regularity conditions assumed in Wang et al. (2014), consistency and asymptotic normality of the estimator $\widehat{\theta}$ can be established and details are omitted. For a point estimator such as $\widehat{\mu}$ defined in (4), its consistency and asymptotic normality can also be established but its asymptotic variance does not have a simple explicit form such as the one for $\widetilde{\mu}$ given in (2). The complication comes from the estimation of μ_0, the correlation between $\widehat{\mu}_0$ and $\bar{y}_k$ in (4), and the estimation of θ that produces $\widehat{\theta}$ correlated with $\widehat{\mu}_0$ and all $\bar{y}_d$'s in (4).

Thus we do not try to obtain an explicit form of the asymptotic variance of $\widehat{\mu}$ defined by (4). Instead, we recommend the bootstrap method for variance estimation or inference. Since our point estimators are all functions of averages and GMM estimators, the general bootstrap theory ensures that the bootstrap variance estimators are consistent and can be effectively applied to avoid the complicated derivation of asymptotic variances of estimators such as $\widehat{\mu}$ in (4), at the expense of a large amount of computations. In Section 3, the performance of bootstrap variance estimators is evaluated by simulations.

2.4. Estimation for multivariate outcomes

In Sections 2.2 and 2.3, we consider the situation where components of $\boldsymbol{y}$ have the same distribution and the population parameter such as the mean μ of a component of $\boldsymbol{y}$ can be estimated using the observed values from all components within each group under assumption (A1) to compensate the missing components. We now consider a multivariate outcome $\boldsymbol{y}$ whose components have different distributions, and we need to estimate population parameters of the jth component y_j of $\boldsymbol{y}$, $j = 1, \ldots, k$. To illustrate, we focus on the estimation of population mean $\mu_j = E(y_j)$ with a fixed $j = 1, \ldots, k$.

To handle the nonignorable nonresponse under assumption (A1), we still group data according to the value of Δ, the number of observed components in $\boldsymbol{y}$, as described in Section 2.1. However, we cannot make use of observations from different components of $\boldsymbol{y}$ within each group; instead, to estimate μ_j we can only use observed values from the fixed jth component. Assuming that $\mu_{j0} = E(y_j \mid \Delta = 0)$ is known, an analog of $\widetilde{\mu}$ in (2) is

$$\widetilde{\mu}_j = \frac{n_0}{n}\mu_{j0} + \sum_{d=1}^{k} \frac{n_d}{n}\bar{y}_{jd}, \tag{8}$$

where $\bar{y}_{jd}$ is the sample mean of observed values of the jth component of $\boldsymbol{y}$ within group $\Delta = d$. The number of observations used for $\bar{y}_{jd}$, n_{jd}, is smaller than the number of observations $dn_d = \sum_{j=1}^{k} n_{jd}$ used for $\bar{y}_d$ in (1). Hence, $\widetilde{\mu}_j$ in (8) may be not stable when the sample size n is not very large. To overcome this difficulty, we consider making use of the always observed covariate $\boldsymbol{x}$ to improve the estimation efficiency.

If a correct parametric model between $\boldsymbol{y}$ and $\boldsymbol{x}$ is imposed, then covariate information can be effectively utilised through the model. Although a linear or parametric relationship between $\boldsymbol{y}$ and $\boldsymbol{x}$ for the whole dataset without nonresponse might be possible, it is unrealistic to expect such relationship still exists between $\boldsymbol{y}$ and $\boldsymbol{x}$ in each group with $\Delta = d$. A purely nonparametric regression between $\boldsymbol{y}$ and $\boldsymbol{x}$ in each group may be applied, but a nonparametric method may be inefficient and suffers from the well-known curse of dimensionality.

A popular approach in sample surveys for improving efficiency without relying on any model between $\boldsymbol{y}$ and $\boldsymbol{x}$ is the Generalised Regression (GREG) method. The GREG is first discussed in Cassel et al. (1976) and studied extensively in the literature; for example, Sarndal et al. (2003) and J. Shao and S. Wang (2014). Since this approach is model-assisted but not model-based, i.e., a model is used to derive efficient estimators that are still asymptotically valid even if the model is incorrect, it suits our purpose of utilising covariates without modelling within each group.

For each d and j, let $\bar{y}_{jd}$ be the sample mean of observed values of the jth component of $\boldsymbol{y}$ within group $\Delta = d$, $\bar{\boldsymbol{x}}_{jd}$ be the sample mean vector of $\boldsymbol{x}$ values corresponding to the observed values used in computing $\bar{y}_{jd}$ within group $\Delta = d$, $\bar{\boldsymbol{x}}_d$ be the sample mean of $\boldsymbol{x}$ values based on all units in group $\Delta = d$, and

$$\widehat{\boldsymbol{\beta}}_{jd} = \left[\sum_{i:\Delta_i = d} r_{ij}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{jd})(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{jd})'\right]^{-1} \times \sum_{i:\Delta_i = d} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{jd}) r_{ij} y_{ij}, \tag{9}$$

which is a least squares estimator based on observed data from jth component of $\boldsymbol{y}$ and $\boldsymbol{x}$ within group $\Delta = d$. Assuming that μ_{j0} is known, our proposed GREG estimator of population mean μ_j is

$$\widetilde{\mu}_j^{GR} = \frac{n_0}{n}\mu_{j0} + \sum_{d=1}^{k}\frac{n_d}{n}\{\bar{y}_{jd} + \widehat{\boldsymbol{\beta}}_{jd}'(\bar{\boldsymbol{x}}_d - \bar{\boldsymbol{x}}_{jd})\}. \quad (10)$$

The following theorem summarises the asymptotic behaviour of the proposed GREG estimator $\widetilde{\mu}_j^{GR}$ in (10), for each fixed $j = 1,\ldots,k$. Note that no model assumption is imposed on the relationship between $\boldsymbol{y}$ and $\boldsymbol{x}$.

Theorem 2.2. *Assume* (A1) *and that, for each* $j = 1,\ldots,k$, *the second-order moments of* $\boldsymbol{x}$ *and* $\boldsymbol{x}y_j$ *are finite, where* y_j *is the jth component of* $\boldsymbol{y}$. *Assume also that, for every* $d = 1,\ldots,k$, $\boldsymbol{\Sigma}_d = \mathrm{Var}(\boldsymbol{x} \mid \Delta = d)$, *the conditional variance of* $\boldsymbol{x}$ *given* $\Delta = d$, *is positive definite. Then, as* $n \to \infty$,

$$\sqrt{n}(\widetilde{\mu}_j^{GR} - \mu_j) \to N\left(0, \tau_j^2 + \sum_{d=0}^{k} p_d\mu_{jd}^2 - \mu_j^2\right)$$

in distribution, (11)

where $\mu_{jd} = E(y_j \mid \Delta = d)$,

$$\tau_j^2 = \frac{1}{n}E\left\{n_\Delta^2\left(\frac{\sigma_{j\Delta}^2}{n_{j\Delta}} - \frac{n_\Delta - n_{j\Delta}}{n_\Delta n_{j\Delta}}\boldsymbol{\beta}_{j\Delta}'\boldsymbol{\Sigma}_\Delta\boldsymbol{\beta}_{j\Delta}\right)\right\}, \quad (12)$$

n_{jd} *is the number of observed* y_{ij}*'s within group* $\Delta = d$, $\sigma_{jd}^2 = \mathrm{Var}(y_j \mid \Delta = d)$, $\boldsymbol{\beta}_{jd} = \boldsymbol{\Sigma}_d^{-1}\mathrm{Cov}(\boldsymbol{x}, y_j \mid \Delta = d)$, $\mathrm{Cov}(\boldsymbol{x}, y_j \mid \Delta = d)$ *is the conditional covariance between* $\boldsymbol{x}$ *and* y_j *given* $\Delta = d$, $d = 1,\ldots,k$, *and* $\boldsymbol{\beta}_{j0}$ *and* σ_{j0}^2 *are defined to be 0. In addition, result* (11) *holds with* $\widetilde{\mu}_j^{GR}$ *replaced by* $\widetilde{\mu}_j$ *in* (8) *and* $\boldsymbol{\beta}_{j\Delta}$ *in* (12) *replaced by* 0.

As indicated by Theorem 2.2, the GREG estimator $\widetilde{\mu}_j^{GR}$ is always asymptotically more efficient than $\widetilde{\mu}_j$ unless $\boldsymbol{\beta}_{dj} = 0$ for all $d = 1,\ldots,k-1$. It can also be seen that $n_d = n_{dj}$ when $d = k$, the group with all completely observed response vectors. This means that the GREG approach does not help in the group $\Delta = k$.

Note that we still need to estimate μ_{j0} for each fixed j. But this can be done using the same approach we discussed in Sections 2.2 and 2.3. Also, the final estimator of μ_j (after replacing μ_{j0} in (10) by its estimator) can be shown to be consistency and asymptotically normal under the same regularity conditions assumed in Wang et al. (2014), but its asymptotic variance does not have a simple explicit form such as the one given in Theorem 2.2. Thus we do not try to obtain an explicit form of the asymptotic variance of the GREG estimator of μ_j. Instead, we recommend the bootstrap method for variance estimation, as we discussed in the end of Section 2.3.

3. Simulation results

In this section, simulation results are presented to investigate the finite sample performance of our proposed estimators developed in Section 2. We consider some different settings. In all simulation studies, the proposed GMM estimator $\widehat{\theta}$ is calculated using the MATLAB function `fminsearch` with initial value $\theta = 0$.

3.1. Results for a single covariate $x = z$ and y with identically distributed components

We first present simulation results under situations where $k = 3$, $\boldsymbol{y} = (y_1, y_2, y_3)$, components y_j's are identically distributed, and there is only a single covariate $\boldsymbol{x} = z$ satisfying (A2), i.e., there is no covariate $\boldsymbol{u}$. Our interest is to estimate the marginal population mean μ of a component of $\boldsymbol{y}$, without applying GREG.

For comparison, in addition to the proposed estimator $\widehat{\mu}$ in (4), we also include the naive estimate $\widehat{\mu}^N$, the sample mean of observed $\boldsymbol{y}$-values, and $\widehat{\mu}^F$, the sample mean when there is no nonresponse, used as a benchmark.

In the first simulation study, z is discrete with $q = 2$ categories, $P(z = 1) = 0.4$ and $P(z = 2) = 0.6$. Conditional on z, $k = 3$ components of $\boldsymbol{y}$ are independently generated from $N(20 + 10z, 8^2)$. Note that components of $\boldsymbol{y}$ are conditionally independent, but are dependent unconditionally, and have the same distribution with unconditional mean $\mu = 36$. Given the generated data, the nonrespondents are generated according to the propensity

$$\pi_\theta(\boldsymbol{y}, z) = [1 + \exp(\alpha + \beta_1y_1 + \beta_2y_2 + \beta_3y_3)]^{-1}, \quad (13)$$

where $\theta = (\alpha, \beta_1, \beta_2, \beta_3)$ with value $(2.5, -0.03, -0.03, -0.03)$ in case I and value $(-3, 0.02, 0.02, 0.02)$ in case II. These values of θ are chosen so that β's have different signs and the unconditional nonresponse probability is approximately between 30% and 40%.

The population in cases III and IV is the same except that z has $q = 3$ categories with $P(z = 1) = 0.3$, $P(z = 2) = 0.3$, and $P(z = 3) = 0.4$, the unconditional population mean is 41, and $\theta = (\alpha, \beta_1, \beta_2, \beta_3) = (2.8, -0.03, -0.03, -0.03)$ and $(-3.3, 0.02, 0.02, 0.02)$ in cases III and IV, respectively.

Table 2 reports simulation results for $n = 2000$ with 1000 simulation runs. The reported quantities are values of estimate, bias in percentage, and standard deviation (SD) for the estimators of μ and parameters in the propensity, based on 1000 simulations. For the estimation of μ, we also calculate the simulation average of the standard error (SE) and coverage probability (CP) of the approximate 95% confidence interval, using the bootstrap variance estimator with bootstrap sample size 100. We do not compute SE and CP for estimators of α

Table 2. Simulation results for a single discrete covariate $\boldsymbol{x} = z$ and $\boldsymbol{y}$ with identically distributed components ($n = 2000$ with 1000 simulations).

Case		Estimation of mean			Estimation in propensity			
		$\widehat{\mu}$	$\widehat{\mu}^N$	$\widehat{\mu}^F$	$\widehat{\alpha}$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
I	Estimate	36.0097	37.3052	36.0029	2.4204	−0.0291	−0.0294	−0.0300
	Bias(%)	0.0269	3.6255	0.0080	−3.1830	−2.9003	−2.0861	0.0493
	SD	0.5927	0.1660	0.1477	0.6110	0.0099	0.0100	0.0099
	SE	0.6197	0.1639	0.1504				
	CP	0.9580	0.0000	0.9520				
II	Estimate	36.0342	35.1975	36.0041	−3.0067	0.0199	0.0202	0.0196
	Bias(%)	0.0951	−2.2291	0.0114	0.2243	−0.2509	0.9542	−1.9291
	SD	0.4321	0.1760	0.1548	0.4128	0.0086	0.0086	0.0087
	SE	0.4675	0.1698	0.1502				
	CP	0.9610	0.0050	0.9470				
III	Estimate	40.9571	43.3364	40.9965	2.7493	−0.0298	−0.0297	−0.0300
	Bias(%)	−0.1046	5.6987	−0.0085	−1.8100	−0.7331	−1.1194	−0.1108
	SD	0.6585	0.2138	0.2068	0.4435	0.0084	0.0088	0.0080
	SE	0.6987	0.2155	0.2113				
	CP	0.9520	0.0000	0.9520				
IV	Estimate	41.0381	39.4074	41.0018	−3.3208	0.0196	0.0202	0.0202
	Bias(%)	0.0930	−3.8843	0.0044	0.6295	−1.9536	1.0658	1.0930
	SD	0.6682	0.2376	0.2125	0.3398	0.0088	0.0080	0.0081
	SE	0.7262	0.2316	0.2119				
	CP	0.9680	0.0000	0.9370				

and β's as parameters in propensity are not the main parameters of interest.

The results in Table 2 show that the GMM estimator $\widehat{\theta}$ and $\widehat{\mu}$ in (4) work well for all cases, in terms of estimation bias, SD, and CP. In addition, the bootstrap SE performs well. The naive estimator $\widehat{\mu}^N$ has a serious positive bias when β's are negative (larger y has smaller nonresponse probability) and has a negative bias when β's are positive (larger y has larger nonresponse probability). Although $\widehat{\mu}^N$ may have a small SD, its bias have a serious effect on inference as its related CP is far from the nominal level 95%.

We next turn to a continuous $z \sim N(0, 4^2)$ and compare different ways to use z in estimation equations in (7). Conditional on z, components of $\boldsymbol{y}$ are independent and identically distributed as $N(30 + 1.5z, 8^2)$, which gives the unconditional mean $\mu = 30$. Given the generated data, the nonrespondents are generated according to (13) with $\theta = (\alpha, \beta_1, \beta_2, \beta_3) = (1.8, -0.03, -0.03, -0.03)$. For the continuous z, we consider three ways of using z in the GMM estimation of θ. In case V, z is discretised into $q = 2$ categories according to the median of z. In case VI, z is discretised into $q = 3$ categories according to the 33% and 66% percentiles of z. In case VII, we use a moment of z, i.e., the vector $\boldsymbol{v}_h$ in (7) has its first two components as $(1, z)$. Results for $n = 2000$ with 1000 simulation runs are given in Table 3, with the same quantities in Table 2.

From the results in Table 3, we can see that cutting z into three categories results in a smaller SD compared with that for discretising z into two categories. Using $(1, z)$ for $\boldsymbol{v}_h$ with a continuous z results in the most efficient estimators of μ among the three ways of using z in (7).

3.2. Results for $\boldsymbol{x} = (u, z)$ and $\boldsymbol{y}$ with identically distributed components

We now add a covariate u into the cases in Section 3.1 and consider $\boldsymbol{x} = (u, z)$ with a univariate continuous u and a categorical z. We consider four cases. In cases VIII–IX, z is a discrete covariate having $q = 2$ categories, $P(z = 1) = 0.4$, and $P(z = 2) = 0.6$. Given z, $u \sim N(10z, 10^2)$. Given $z = 1$ and u, components of $\boldsymbol{y}$ are independent and identically distributed as $N(u + 5z, 8^2)$; given $z = 2$ and u, components of $\boldsymbol{y}$ are independent and identically distributed as $N(10 + 0.5u + 5z, 8^2)$. The unconditional mean μ is 24. The propensity is

$$\pi_\theta(\boldsymbol{y}, u) = [1 + \exp(\alpha + \beta_1 y_1 + \beta_2 y_2 + \beta_3 y_3 + \gamma u)]^{-1}, \tag{14}$$

where $\theta = (\alpha, \beta_1, \beta_2, \beta_3, \gamma) = (0.6, -0.03, -0.03, -0.03, 0.04)$ in case VIII and $(1.7, -0.03, -0.03, -0.03, -0.04)$ in case IX. These values are chosen so that γ has different signs and the unconditional nonresponse probability is approximately between 30% and 40%.

In cases X–XI, z has $q = 3$ categories, $P(z = 1) = 0.3$, $P(z = 2) = 0.3$ and $P(z = 3) = 0.4$. Given z, $u \sim N(10z, 10^2)$. Given $z = 1$ and u, components of $\boldsymbol{y}$ are independent and identically distributed as $N(u + 5z, 3^2)$; given $z = 2$ and u, components of $\boldsymbol{y}$ are independent and identically distributed as $N(1.5u + 5z, 3^2)$; given $z = 3$ and u, components of $\boldsymbol{y}$ are independent and identically distributed as $N(10 + 0.5u + 5z, 3^2)$. The unconditional mean μ is 32.5. The propensity is given by (14) with $\theta = (\alpha, \beta_1, \beta_2, \beta_3, \gamma) = (1, -0.03, -0.03, -0.03, 0.05)$ in case X and $(2.8, -0.03, -0.03, -0.03, -0.05)$ in case XI.

Table 3. Simulation results for a single continuous covariate $\boldsymbol{x} = z$ and $\boldsymbol{y}$ with identically distributed components ($n = 2000$ with 1000 simulations).

Case		Estimation of mean			Estimation in propensity			
		$\widehat{\mu}$	$\widehat{\mu}^N$	$\widehat{\mu}^F$	$\widehat{\alpha}$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
V	Estimate	30.0268	31.4417	29.9946	1.7769	−0.0304	−0.0302	−0.0298
	Bias(%)	0.0893	4.8056	−0.0180	−1.2853	1.4452	0.5892	−0.6491
	SD	0.3932	0.1884	0.1767	0.4663	0.0114	0.0111	0.0115
	SE	0.4333	0.1814	0.1687				
	CP	0.9770	0.0000	0.9390				
VI	Estimate	30.0194	31.4549	30.0098	1.7658	−0.0300	−0.0298	−0.0301
	Bias(%)	0.0647	4.8497	0.0327	−1.8974	0.0565	−0.7405	0.3017
	SD	0.3697	0.1805	0.1659	0.4278	0.0102	0.0093	0.0102
	SE	0.3918	0.1810	0.1686				
	CP	0.9620	0.0000	0.9470				
VII	Estimate	30.0441	31.4541	30.0040	1.7513	−0.0292	−0.0274	−0.0373
	Bias(%)	0.1471	4.8470	0.0135	−2.7030	−2.7809	−8.6993	24.4380
	SD	0.3511	0.1828	0.1676	2.3737	0.0270	0.1044	0.2069
	SE	0.3926	0.1810	0.1686				
	CP	0.9720	0.0000	0.9490				

Results for $n = 2000$ with 1000 simulation runs are given in Table 4. Conclusions for results in Table 4 are similar to those in Tables 2 and 3.

3.3. Results for a multivariate outcome y

In this section, we present simulation results under situations where $k = 3$, components of $\boldsymbol{y} = (y_1, y_2, y_3)$ have different distributions, and our interest is to estimate each marginal population mean $\mu_j = E(y_j)$, $j = 1, \ldots, k$. We consider the proposed GREG estimator $\widehat{\mu}_j^{GR}$ as well as the estimator $\widehat{\mu}_j$ without applying GREG, $j = 1, \ldots, k$. The naive estimator $\widehat{\mu}_j^N$, the sample mean of observed values of y_j, and $\widehat{\mu}_j^F$, the sample mean of y_j when there is no nonresponse, are also included.

We consider $\boldsymbol{x} = (u, z)$ with independent u and z, where u is continuous and distributed as $N(3, 5^2)$. In cases XII–XIII, z is continuous and distributed as $N(2, 1)$ and given z and u, $y_1 \sim N(u + 3z, 3^2)$, $y_2 \sim N(u + 4z, 3^2)$, $y_3 \sim N(2u + 5z, 3^2)$ and y_j's are independent. The unconditional mean vector $E(\boldsymbol{y})$ is $(9, 11, 16)$. In cases XIV–XV, z is discrete with $q = 3$ categories, $P(z = 1) = 0.3$, $P(z = 2) = 0.3$ and $P(z = 3) = 0.4$; given z and u, $y_1 \sim N(2u + 2z, 3^2)$, $y_2 \sim N(2u + 4z, 3^2)$, $y_3 \sim N(4u + 2z, 3^2)$, and y_j's are independent. The unconditional mean μ is $(10.2, 14.4, 16.2)$. The propensity is given by (14) with $\theta = (\alpha, \beta_1, \beta_2, \beta_3, \gamma) = (0.1, -0.02, -0.02, -0.02, 0.05)$ in cases XII and XIV and $(-1.2, 0.02, 0.02, 0.02, -0.1)$ in cases XIII and XV. These values are chosen so that γ has different signs and the unconditional nonresponse probability is approximately between 30% and 40%.

Results for $n = 2000$ with 1000 simulation runs are given in Table 5. The results show that both proposed estimators $\widehat{\mu}_j$ and $\widehat{\mu}_j^{GR}$ perform well for each component of $\boldsymbol{y}$ under all cases with coverage probabilities close to the nominal level 0.95. They are much better compared with the naive biased estimator $\widehat{\mu}_j^N$. Also, the estimator $\widehat{\mu}_j^{GR}$ with GREG has a respectable improvement in standard deviation compared with $\widehat{\mu}_j$ without GREG.

4. Real data examples

We apply our proposed estimators to two real data sets from the National Longitudinal Survey of Mature and Young Women (NLSW) and the National Health and Nutrition Examination Survey (NHANES). The proposed estimation approach introduced in Section 2.2 is applied on the NLSW survey data since components of the outcome we choose from the dataset can be treated as from the same distribution. The proposed estimation method introduced in Section 2.4 with or without the GREG is applied on the NHANES data since the outcome we choose from the dataset is multivariate.

We present the estimated values and standard error (SE) under bootstrap method of the marginal means as well as the estimated values of the parameters in the nonresponse propensity. Our results and conclusions are based on assumptions (A1)–(A2) which, unfortunately, cannot be checked using available data. The assumption that components of $\boldsymbol{r}$ are conditionally independent and identically distributed given $(\boldsymbol{y}, \boldsymbol{x})$ seems reasonable from the specific problems under investigation.

4.1. Application to NLSW data

The NLSW started in the mid-1960s because the U.S. Department of Labor was interested in studying the employment patterns of non-institutionalised civilian women in the United States. We focus on the survey of mature women cohort with ages from 30s to early 40s. A detailed description of this survey can be found at https://www.bls.gov/nls/original-cohorts/mature-and-young-women.htm.

Among many topics, we focus on the variable of women's weight in pounds (ERNYR-P) from heath topic as our example. More specifically, we consider the outcome $\boldsymbol{y} = (y_1, y_2, y_3)$ $(k = 3)$, where y_j's are weights

Table 4. Simulation results for $\boldsymbol{x} = (u, z)$ with a categorical z and a continuous u ($n = 2000$ with 1000 simulations).

Case			Estimation of mean			
Case			$\widehat{\mu}$	$\widehat{\mu}_N$		$\widehat{\mu}_F$
VIII	Estimate		24.0085	26.2166		23.9988
	Bias(%)		0.0354	9.2358		−0.0052
	SD		0.3526	0.2484		0.2627
	SE		0.3763	0.2406		0.2543
	CP		0.9620	0.0000		0.9450
		Estimation in propensity				
		$\widehat{\alpha}$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\eta}$
	Estimate	0.5253	−0.0290	−0.0297	−0.0292	0.0384
	Bias(%)	−12.4446	−3.4645	−1.0577	−2.7879	−3.9445
	SD	0.2599	0.0122	0.0122	0.0124	0.0152
			Estimation of mean			
			$\widehat{\mu}$	$\widehat{\mu}_N$		$\widehat{\mu}_F$
VIV	Estimate		24.0020	27.8904		24.0086
	Bias(%)		0.0083	16.2098		0.0358
	SD		0.4641	0.2085		0.2479
	SE		0.4623	0.2118		0.2541
	CP		0.9780	0.0000		0.9500
		Estimation in propensity				
		$\widehat{\alpha}$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\eta}$
	Estimate	1.2077	−0.0326	−0.0270	−0.0290	−0.0468
	Bias(%)	−28.9598	8.7831	−9.9284	−3.4798	17.0849
	SD	2.4568	0.4101	0.0505	0.2869	0.1186
			Estimation of mean			
Case			$\widehat{\mu}$	$\widehat{\mu}_N$		$\widehat{\mu}_F$
X	Estimate		32.5066	35.3083		32.5009
	Bias(%)		0.0203	8.6409		0.0028
	SD		0.3980	0.2608		0.2785
	SE		0.4354	0.2596		0.2828
	CP		0.9690	0.0000		0.9570
		Estimation in propensity				
		$\widehat{\alpha}$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\eta}$
	Estimate	0.9710	−0.0300	−0.0295	−0.0306	0.0499
	Bias(%)	−2.8962	−0.0200	−1.6826	2.0488	−0.2297
	SD	0.2568	0.0273	0.0257	0.0268	0.0119
			Estimation of mean			
			$\widehat{\mu}$	$\widehat{\mu}_N$		$\widehat{\mu}_F$
XI	Estimate		32.3482	37.5787		32.5173
	Bias(%)		−0.4670	15.6268		0.0532
	SD		1.0728	0.2053		0.2869
	SE		1.7310	0.1811		0.2661
	CP		0.9400	0.0000		0.9500
		Estimation in propensity				
		$\widehat{\alpha}$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\eta}$
	Estimate	2.6765	−0.0276	−0.0283	−0.0332	−0.0533
	Bias(%)	−4.4104	−8.1382	−5.5310	10.6065	6.6456
	SD	0.6086	0.0808	0.0672	0.0743	0.0224

(in lbs) of respondent in 1997, 1999 and 2001, respectively. The outcome values are self-reported in roughly every 2 years. Since the participants are matured women, the three components of $\boldsymbol{y}$ have almost the same distribution. We are interested in estimating the overall population mean μ of the weight using the proposed method in Section 2.2. We use the age of participant when she joined the NLSW survey as the nonresponse instrument z.

In the dataset, each of three components of $\boldsymbol{y}$ has about 29% nonresponse probability while the covariate has no nonresponse. The number of observed values in

Table 5. Simulation results for $\boldsymbol{x} = (u, z)$ and multivariate $\boldsymbol{y}$ ($n = 2000$ with 1000 simulations); SDimp(%) = (1− SD of $\hat{\mu}_j^{GR}$/ SD of $\hat{\mu}_j$) × 100%.

Case				Estimation of mean			
Case				$\widehat{\mu}_j$	$\widehat{\mu}_j^{GR}$	$\widehat{\mu}_j^{N}$	$\widehat{\mu}_j^{F}$
XII	$j = 1$	Estimate		9.0131	9.0137	9.6294	9.0064
		Bias(%)		0.1456	0.1522	6.9935	0.0715
		SD		0.2116	0.1843	0.1856	0.1469
		SE		0.2179	0.1897	0.1851	0.1463
		CP		0.9550	0.9470	0.0920	0.9540
		SDimp(%)			12.8762		
	$j = 2$	Estimate		11.0105	11.0204	11.7195	11.0076
		Bias(%)		0.0959	0.1852	6.5413	0.0687
		SD		0.2327	0.2043	0.1971	0.1590
		SE		0.2377	0.2064	0.2001	0.1580
		CP		0.9590	0.9580	0.0650	0.9370
		SDimp(%)			12.1987		
	$j = 3$	Estimate		11.0105	16.0257	17.1169	16.0077
		Bias(%)		0.2687	0.1608	6.9809	0.0484
		SD		0.3810	0.3146	0.3308	0.2624
		SE		0.3819	0.3188	0.3272	0.2580
		CP		0.9510	0.9570	0.0900	0.9400
		SDimp(%)			17.4088		
			Estimation in propensity				
			$\widehat{\alpha}$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\gamma}$
		Estimate	0.1036	−0.0201	−0.0206	−0.0208	0.0516
		Bias(%)	3.5808	0.7052	3.0718	4.2204	3.1186
		SD	0.1584	0.0251	0.0210	0.0186	0.0263
				Estimation of mean			
Case				$\widehat{\mu}_j$	$\widehat{\mu}_j^{GR}$	$\widehat{\mu}_j^{N}$	$\widehat{\mu}_j^{F}$
XIII	$j = 1$	Estimate		9.0092	9.0094	8.8764	9.0034
		Bias(%)		0.1020	0.1040	−1.3734	0.0381
		SD		0.2116	0.1830	0.1784	0.1470
		SE		0.2124	0.1865	0.1769	0.1459
		CP		0.9560	0.9530	0.9050	0.9450
		SDimp(%)			13.5115		
	$j = 2$	Estimate		11.0071	11.0110	10.7956	11.0005
		Bias(%)		0.0642	0.1004	−1.8585	0.0046
		SD		0.2283	0.2002	0.1836	0.1532
		SE		0.2328	0.2044	0.1904	0.1567
		CP		0.9430	0.9460	0.8150	0.9430
		SDimp(%)			12.3331		
	$j = 3$	Estimate		11.0071	16.0159	15.8899	16.0062
		Bias(%)		0.1387	0.0994	−0.6879	0.0390
		SD		0.3563	0.3071	0.3090	0.2535
		SE		0.3663	0.3084	0.3118	0.2570
		CP		0.9400	0.9630	0.9230	0.9410
		SDimp(%)			13.8014		
			Estimation in propensity				
			$\widehat{\alpha}$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\gamma}$
		Estimate	−1.2133	0.0186	0.0204	0.0209	−0.1014
		Bias(%)	1.1097	−7.0930	1.8714	4.4117	1.4445
		SD	0.1326	0.0255	0.0197	0.0166	0.0247
				Estimation of mean			
Case				$\widehat{\mu}_j$	$\widehat{\mu}_j^{GR}$	$\widehat{\mu}_j^{N}$	$\widehat{\mu}_j^{F}$
XIV	$j = 1$	Estimate		10.2059	10.2035	12.2317	10.1935
		Bias(%)		0.0579	0.0345	19.9190	−0.0638
		SD		0.3286	0.2885	0.2803	0.2353
		SE		0.3416	0.2953	0.2880	0.2354
		CP		0.9560	0.9550	0.0000	0.9520
		SDimp(%)			12.1951		

(*continued*).

Table 5. Continued.

Case			Estimation of mean			
			$\widehat{\mu}_j$	$\widehat{\mu}_j^{GR}$	$\widehat{\mu}_j^{N}$	$\widehat{\mu}_j^{F}$
	$j=2$	Estimate	14.4110	14.4100	16.5011	14.3915
		Bias(%)	0.0764	0.0698	14.5906	−0.0588
		SD	0.3461	0.2952	0.2964	0.2436
		SE	0.3459	0.2966	0.2986	0.2439
		CP	0.9550	0.9500	0.0000	0.9520
		SDimp(%)		14.6967		
	$j=3$	Estimate	14.4110	16.2097	20.1238	16.1886
		Bias(%)	0.0064	0.0601	24.2210	−0.0705
		SD	0.6786	0.5654	0.5603	0.4542
		SE	0.6633	0.5645	0.5535	0.4517
		CP	0.9490	0.9530	0.0000	0.9410
		SDimp(%)		16.6760		

	Estimation in propensity				
	$\widehat{\alpha}$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\gamma}$
Estimate	0.0853	−0.0209	−0.0206	−0.0189	0.0466
Bias(%)	−14.7282	4.4709	2.8384	−5.4701	−6.8876
SD	0.1920	0.0305	0.0186	0.0303	0.1200

Case			Estimation of mean			
			$\widehat{\mu}_j$	$\widehat{\mu}_j^{GR}$	$\widehat{\mu}_j^{N}$	$\widehat{\mu}_j^{F}$
XV	$j=1$	Estimate	10.2144	10.2160	9.0712	10.2000
		Bias(%)	0.1410	0.1566	−11.0664	0.0004
		SD	0.3727	0.3264	0.2830	0.2346
		SE	0.3870	0.3436	0.2867	0.2355
		CP	0.9650	0.9620	0.0260	0.9450
		SDimp(%)		12.4127		
	$j=2$	Estimate	14.4282	14.4221	13.2057	
		Bias(%)	0.1961	0.1535	−8.2938	0.0095
		SD	0.4093	0.3620	0.3017	0.2433
		SE	0.4235	0.3799	0.2965	0.2436
		CP	0.9660	0.9610	0.0230	0.9400
		SDimp(%)		11.5561		
	$j=3$	Estimate	14.4282	16.2286	14.0698	16.1996
		Bias(%)	0.1344	0.1768	−13.1493	−0.0026
		SD	0.6812	0.5890	0.5366	0.4531
		SE	0.7135	0.6175	0.5504	0.4518
		CP	0.9600	0.9600	0.0400	0.9470
		SDimp(%)		13.5300		

	Estimation in propensity				
	$\widehat{\alpha}$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\gamma}$
Estimate	−1.5441	0.0105	0.0267	0.0240	−0.1351
Bias(%)	28.6770	−47.6082	33.4260	19.9365	35.0864
SD	9.8790	0.2753	0.1834	0.1707	1.3114

each nonresponse pattern for the outcome y is shown in Table 6.

We computed the proposed estimator $\widehat{\mu}$ in Sections 2.2 and 2.3. Since the covariate $x =$ 'age of respondent when joining the survey' is univariate and continuous, we treat $x = z$ and use the moments of z directly in the GMM algorithm. The results are given in Table 7 and the SE is computed as the squared root of the bootstrap variance estimate with bootstrap size 100.

For comparison, we include the naive estimator $\widehat{\mu}^N$, the sample mean of observed y values. We can see that our proposed estimator $\widehat{\mu}$ has a significant difference from the naive estimate $\widehat{\mu}^N$.

Table 7. Estimation based on NLSW data.

ERNYR-P		Estimation of marginal population mean			
		$\widehat{\mu}$		$\widehat{\mu}^N$	
	Estimate	132.2044		158.1434	
	SE	1.2182		0.6510	
		Estimation of parameters in propensity			
Parameter		$\widehat{\alpha}$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
	Estimate	−0.8844	−0.0245	0.0489	−0.0374

4.2. Application to NHANES data

The NHANES is a major program of the National Center for Health Statistics, which is a part of the Centers for Disease Control and Prevention responsible for producing vital and health statistics for the United States. The NHANES is a program designed to assess the health and nutritional status of adults and children in the non-institutionalised civilian resident population of the United States. A description of this survey can be found at https://www.cdc.gov/nchs/nhanes/about_nhanes.htm.

The NHANES program began in the early 1960s and had been conducted as a series of surveys focusing on different population groups or health topics. In 1999, the survey became a continuous program that has a changing focus on a variety of health and nutrition measurements to meet emerging needs. The survey is unique in that it combines interviews and physical examinations. The home-interview part collects answers from demographic, socioeconomic, dietary, and health-related questions. The examination component conducted in a mobile examination centre consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel.

Table 6. The number of observed values in each nonresponse pattern

Group	Nonresponse pattern			Number of observations
	Weight in 1997	Weight in 1999	Weight in 2001	
$\Delta=0$	?	?	?	504
$\Delta=1$	✓	?	?	119
	?	✓	?	67
	?	?	✓	90
$\Delta=2$	✓	✓	?	158
	✓	?	✓	134
	?	✓	✓	143
$\Delta=3$	✓	✓	✓	1799
Total				3014

Table 8. The number of observed values in each of nonresponse pattern.

	Missing pattern			
Group	LBXSCH	BPXSY1	BMDAVSAD	Number of observations
$\Delta = 0$	?	?	?	2007
$\Delta = 1$	✓	?	?	94
	?	✓	?	91
	?	?	✓	90
$\Delta = 2$	✓	✓	?	374
	✓	?	✓	422
	?	✓	✓	444
$\Delta = 3$	✓	✓	✓	5578
Total				9100

The data set we focused on is for 2013–2014 consisting of 9100 persons who completed both interview and examination. We consider a multivariate outcome $\boldsymbol{y}$ with $k = 3$ and two demographic covariates from the dataset. The three components of $\boldsymbol{y}$ are the total cholesterol (mg/dL), 'LBXSCH', the first reading of systolic blood pressure (mm Hg), 'BPXSY1', and the average sagittal abdominal diameter (cm), 'BMDAVSAD'. The two covariates are the age in years of the household reference person, 'DMDHRAGE', and the total household income (reported as a range value in dollars), 'INDHHIN2'.

Each of the three components of $\boldsymbol{y}$ has about 28% missing values while two covariates have no missing value. The number of observed values in each of nonresponse pattern for $\boldsymbol{y}$ is shown in Table 8.

In this example, the three components of $\boldsymbol{y}$ have different distributions and we are interested in estimating the population mean for each y_j. Therefore, we apply our proposed estimator in Section 2.4 with GREG, denoted by $\widehat{\mu}_j^{GR}$, and the estimator without generalised regression, denoted by $\widehat{\mu}_j$. For comparison, we also include the naive estimate $\widehat{\mu}_j^N$, the sample mean of observed y_j values.

Since $\boldsymbol{x}$ is two dimensional, we try two scenarios, $z =$ DMDHRAGE and $u =$ INDHHIN2 in case 1, and $z =$ INDHHIN2 and $u =$ DMDHRAGE in case 2. The propensity model we used is given by (14).

The results for two cases are given in Table 9, where SE is computed as the squared root of the bootstrap variance estimate with bootstrap size 100.

From both cases, we can see that estimators $\widehat{\mu}_j$ and $\widehat{\mu}_j^{GR}$ are very similar but are significantly different from the naive estimator $\widehat{\mu}_j^N$, indicating that the naive estimator is biased according to our theory and empirical results. The fact that different ways of defining z in (A2) result in very similar estimates of μ_j's indicates that both covariates DMDHRAGE and INDHHIN2 are suitable to be used as z in (A2), although different z's produce different estimates of parameters in propensity. In this example, covariates may not help very much in estimating the marginal population means, although they are very helpful in handling nonignorable nonresponse.

Table 9. Estimation based on NHANES data.

Results for $z =$ DMDHRAGE

		Estimation of mean		
Component		$\widehat{\mu}_j$	$\widehat{\mu}_j^{GR}$	$\widehat{\mu}_j^N$
LBXSCH	Estimate	144.0550	144.0248	184.5257
	SE	0.8946	0.9050	0.4634
BPXSY1	Estimate	89.1456	89.2445	119.2604
	SE	16.6209	16.6619	21.5088
BMDAVSAD	Estimate	0.5489	0.5500	0.2094
	SE	0.1059	0.1066	0.0560

	Estimation in propensity				
	$\widehat{\alpha}$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\gamma}$
Estimate	0.1740	0.0066	−0.0159	−0.1152	−0.0183
SE	2.4623	0.0105	0.0134	0.1161	0.0113

Results for $z =$ INDHHIN2

		Estimation of mean		
Component		$\widehat{\mu}_j$	$\widehat{\mu}_j^{GR}$	$\widehat{\mu}_j^N$
LBXSCH	Estimate	144.0251	143.9984	184.5257
	SE	0.8394	0.8407	0.5822
BPXSY1	Estimate	89.0743	89.1804	119.2758
	SE	0.5200	0.5169	0.2308
BMDAVSAD	Estimate	16.6357	16.6790	21.5052
	SE	0.1020	0.1038	0.0658

	Estimation in propensity				
	$\widehat{\alpha}$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\gamma}$
Estimate	−1.5423	−0.0035	0.0055	−0.0978	0.0090
SE	6.2862	0.0346	0.0205	0.2393	0.0105

5. Technical proofs

Proof of Theorem 2.1: The asymptotic normality result (3) follows from the Central Limit Theorem. Hence, it remains to show that the asymptotic mean and variance are of the given form. Let $\boldsymbol{\Delta} = \{\Delta_i, \ldots, \Delta_i, i = 1, \ldots, n\}$. From conditioning,

$$E(\widetilde{\mu}) = E\left\{E\left(\frac{n_0}{n}\mu_0 + \sum_{d=1}^{k}\frac{n_d}{n}\bar{y}_d \,\middle|\, \boldsymbol{\Delta}\right)\right\}$$
$$= E\left\{\sum_{d=0}^{k}\frac{n_d}{n}\mu_d\right\} = \sum_{d=0}^{k}p_d\mu_d = \mu,$$

so that the mean of $\widetilde{\mu} - \mu$ is 0. To derive the asymptotic variance, we calculate

$$\mathrm{Var}(\widetilde{\mu}) = \mathrm{Var}\left(\frac{n_0}{n}\mu_0 + \sum_{d=1}^{k}\frac{n_d}{n}\bar{y}_d\right)$$
$$= E\left\{\mathrm{Var}\left(\sum_{d=1}^{k}\frac{n_d}{n}\bar{y}_d \,\middle|\, \boldsymbol{\Delta}\right)\right\}$$
$$+ \mathrm{Var}\left\{E\left(\frac{n_0}{n}\mu_0 + \sum_{d=1}^{k}\frac{n_d}{n}\bar{y}_d \,\middle|\, \boldsymbol{\Delta}\right)\right\}$$

$$
\begin{aligned}
&= E\left\{\frac{1}{n^2}\sum_{d=1}^{k} n_d\sigma_d^2\right\} + \mathrm{Var}\left\{\sum_{d=0}^{k}\frac{n_d}{n}\mu_d\right\} \\
&= \frac{1}{n}\sum_{d=1}^{k} p_d\sigma_d^2 + \frac{1}{n^2}\sum_{d=0}^{k}\mu_d^2\mathrm{Var}(n_d) \\
&\quad + \frac{1}{n^2}\sum_{d\neq l}\mu_d\mu_l\mathrm{Cov}(n_d, n_l) \\
&= \frac{1}{n}\sum_{d=0}^{k} p_d\sigma_d^2 + \frac{1}{n}\sum_{d=0}^{k}\mu_d^2 p_d(1-p_d) \\
&\quad - \frac{1}{n}\sum_{d\neq l}\mu_d\mu_l p_d p_l,
\end{aligned}
$$

where the last equality follows from the fact that the vector $(n_0, n_1, \ldots, n_k)$ follows a multinomial distribution so that $\mathrm{Var}(n_d) = np_d(1-p_d)$ and $\mathrm{Cov}(n_d, n_l) = -np_dp_l$ for any $d \neq l$. Then, the result follows from

$$
\sum_{d=0}^{k}\mu_d^2p_d^2 + \sum_{d\neq l}\mu_d\mu_l p_d p_l = \left(\sum_{d=0}^{k} p_d\mu_d\right)^2 = \mu^2.
$$

■

Lemma 5.1. *Under the conditions of Theorem 2.2, for each $j = 1, \ldots, k$ and each $d = 1, \ldots, k$, $\widehat{\boldsymbol{\beta}}_{jd} \to \boldsymbol{\beta}_{jd}$ in probability as $n \to \infty$, where $\widehat{\boldsymbol{\beta}}_{jd}$ is defined in (9) and $\boldsymbol{\beta}_{jd} = \boldsymbol{\Sigma}_d^{-1}\mathrm{Cov}(\boldsymbol{x}, y_j \mid \Delta = d)$.*

Proof of Lemma 5.1: For fixed j and d, by (A1), the weak law of large numbers for independent random variables, and Lemma 2.1 in Section 2.1, as $n \to \infty$,

$$
\begin{aligned}
&\frac{1}{n_{jd}}\sum_{i:\Delta_i=d} r_{ij}\boldsymbol{x}_i y_{ij} \to E(\boldsymbol{x}y_j \mid \Delta = d), \\
&\bar{\boldsymbol{x}}_{jd} \to E(\boldsymbol{x} \mid \Delta = d), \quad \bar{y}_{jd} \to E(y_j \mid \Delta = d)
\end{aligned}
$$

in probability. Therefore, as $n \to \infty$,

$$
\begin{aligned}
&\frac{1}{n_{jd}}\sum_{i:\Delta_i=d}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{jd})r_{ij}y_{ij} \\
&\to E(\boldsymbol{x}y_j \mid \Delta = d) - E(\boldsymbol{x} \mid \Delta = d)E(y_j \mid \Delta = d) \\
&= \mathrm{Cov}(\boldsymbol{x}, y_j \mid \Delta = d) \text{ in probability.}
\end{aligned}
$$

Similarly, it can be shown that

$$
\frac{1}{n_{jd}}\sum_{i:\Delta_i=d} r_{ij}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_i)' \to \boldsymbol{\Sigma}_d \text{ in probability.}
$$

The proof is completed by combining the results and using the definitions of $\widehat{\boldsymbol{\beta}}_{jd}$ and $\boldsymbol{\beta}_{jd}$. ■

Proof of Theorem 2.2: From (10),

$$
\begin{aligned}
\widetilde{\mu}_j^{GR} - \mu_j &= \frac{n_0}{n}\mu_{j0} + \sum_{d=1}^{k}\frac{n_d}{n}[\bar{y}_{jd} + \widehat{\boldsymbol{\beta}}_{jd}'(\bar{\boldsymbol{x}}_d - \bar{\boldsymbol{x}}_{jd})] \\
&\quad - \sum_{d=0}^{k} p_d\mu_{jd} \\
&= \sum_{d=0}^{k}\left(\frac{n_d}{n} - p_d\right)\mu_{jd} \\
&\quad + \sum_{d=1}^{k}\frac{n_d}{n}[(\bar{y}_{jd} - \mu_{jd}) + \widehat{\boldsymbol{\beta}}_{jd}'(\bar{\boldsymbol{x}}_d - \bar{\boldsymbol{x}}_{jd})] \\
&= U_j + V_j + W_j,
\end{aligned}
$$

where

$$
\begin{aligned}
U_j &= \sum_{d=1}^{k}\frac{n_d}{n}[(\bar{y}_{jd} - \mu_{jd}) + \boldsymbol{\beta}_{jd}'(\bar{\boldsymbol{x}}_d - \bar{\boldsymbol{x}}_{jd})], \\
V_j &= \sum_{d=0}^{k}\left(\frac{n_d}{n} - p_d\right)\mu_{jd}, \\
W_j &= \sum_{d=1}^{k}\frac{n_d}{n}[(\widehat{\boldsymbol{\beta}}_{jd} - \boldsymbol{\beta}_{jd})'(\bar{\boldsymbol{x}}_d - \bar{\boldsymbol{x}}_{jd})].
\end{aligned}
$$

By Lemma 5.1, W_j is asymptotically negligible compared with U_j and V_j. Hence, to prove (11), it suffices to show that $\sqrt{n}(U_j + V_j)$ converges in distribution to the limiting normal distribution in (11). Consider V_j first. Note that

$$
\begin{aligned}
V_j &= \sum_{d=0}^{k}\frac{n_d}{n}\mu_{jd} - \mu_j = \frac{1}{n}\sum_{d=0}^{k}\sum_{i:\Delta_i=d}\mu_{jd} - \mu_j \\
&= \frac{1}{n}\sum_{i:\Delta_i=d}\sum_{d=0}^{k}\mu_{jd} - \mu_j.
\end{aligned}
$$

Then

$$
E(V_j) = 0 \quad \text{and} \quad \mathrm{Var}(V_j) = \frac{1}{n}\left(\sum_{d=0}^{k} p_d\mu_{jd}^2 - \mu_j^2\right).
$$

From the Central Limit Theorem,

$$
\frac{\sqrt{n}V_j}{\sqrt{\mathrm{Var}(V_j)}} \to N(0,1) \text{ in distribution.}
$$

We now turn to U_j. Let $\boldsymbol{\Delta} = \{\Delta_1, \ldots, \Delta_n\}$. Conditioned on $\boldsymbol{\Delta}$,

$$
\begin{aligned}
&E(U_j \mid \boldsymbol{\Delta}) \\
&= E\left\{\sum_{d=1}^{k}\frac{n_d}{n}[(\bar{y}_{jd} - \mu_{jd}) + \boldsymbol{\beta}_{jd}'(\bar{\boldsymbol{x}}_d - \bar{\boldsymbol{x}}_{jd})] \,\middle|\, \boldsymbol{\Delta}\right\}
\end{aligned}
$$

$$= \sum_{d=1}^{k} \frac{n_d}{n} E\left(\bar{y}_{jd} - \mu_{jd} \,\middle|\, \mathbf{\Delta} \right)$$
$$+ \sum_{d=1}^{k} \frac{n_d}{n} \boldsymbol{\beta}'_{jd} E\left(\bar{\boldsymbol{x}}_d - \bar{\boldsymbol{x}}_{jd} \,\middle|\, \mathbf{\Delta} \right)$$
$$= 0,$$

where the last equality follows from $E(\bar{y}_{jd} - \mu_{jd} \mid \mathbf{\Delta}) = 0$ and $E(\bar{\boldsymbol{x}}_d - \bar{\boldsymbol{x}}_{jd} \mid \mathbf{\Delta}) = 0$ as given $\mathbf{\Delta}$, $\boldsymbol{x}$ and y_j values in group $\Delta = d$ are exchangeable. It follows from the Central Limit Theorem that, conditioned on $\mathbf{\Delta}$,

$$\frac{\sqrt{n} U_j}{\sqrt{\mathrm{Var}(U_j \mid \mathbf{\Delta})}} \to N(0,1) \text{ in distribution.}$$

Then, unconditionally,

$$\frac{\sqrt{n} U_j}{\sqrt{E\{\mathrm{Var}(U_j \mid \mathbf{\Delta})\}}} \to N(0,1) \text{ in distribution.}$$

To complete the proof, it remains to show two items. One is $n^{-1}E\{\mathrm{Var}(U_j \mid \mathbf{\Delta})\} = \tau_j^2$ given in (12); the other is that $\mathrm{Cov}(U_j, V_j) = 0$. The latter follows from

$$\begin{aligned}\mathrm{Cov}(U_j, V_j) &= \mathrm{Cov}\{E(U_j \mid \mathbf{\Delta}), E(V_j \mid \mathbf{\Delta})\} \\ &\quad + E\{\mathrm{Cov}(U_j, V_j \mid \mathbf{\Delta})\} \\ &= \mathrm{Cov}\{0, E(V_j \mid \mathbf{\Delta})\} + 0 \\ &= 0,\end{aligned}$$

where the second equality follows from the fact that V_j is a function of $\mathbf{\Delta}$ so that $\mathrm{Cov}(U_j, V_j \mid \mathbf{\Delta}) = 0$ almost surely. To calculate $E\{\mathrm{Var}(U_j \mid \mathbf{\Delta})\}$, note that, for any fixed d,

$$\bar{y}_{jd} + \boldsymbol{\beta}'_{jd}(\bar{\boldsymbol{x}}_d - \bar{\boldsymbol{x}}_{jd}) = \bar{y}_{jd} + \frac{n_d - n_{jd}}{n_d} \boldsymbol{\beta}'_{jd}(\tilde{\boldsymbol{x}}_{jd} - \bar{\boldsymbol{x}}_{jd}),$$

where

$$\bar{\boldsymbol{x}}_{jd} = \frac{1}{n_{jd}} \sum_{i:\Delta_i = d} r_{ij} \boldsymbol{x}_i,$$
$$\tilde{\boldsymbol{x}}_{jd} = \frac{1}{n_d - n_{jd}} \sum_{i:\Delta_i = d} (1 - r_{ij}) \boldsymbol{x}_i.$$

Since observations in $\bar{\boldsymbol{x}}_{jd}$ and $\tilde{\boldsymbol{x}}_{jd}$ are not overlapped, conditioned on $\mathbf{\Delta}$,

$$\mathrm{Var}\left(\bar{y}_{jd} + \boldsymbol{\beta}'_{jd}(\bar{\boldsymbol{x}}_d - \bar{\boldsymbol{x}}_{jd}) \mid \mathbf{\Delta} \right)$$
$$= \mathrm{Var}\left\{ \bar{y}_{jd} + \frac{n_d - n_{jd}}{n_d} \boldsymbol{\beta}'_{jd}(\tilde{\boldsymbol{x}}_{jd} - \bar{\boldsymbol{x}}_{jd}) \,\middle|\, \mathbf{\Delta} \right\}$$
$$= \mathrm{Var}\left(\bar{y}_{jd} \mid \mathbf{\Delta} \right)$$
$$+ \frac{(n_d - n_{jd})^2}{n_d^2} \boldsymbol{\beta}'_{jd} \mathrm{Var}\left(\tilde{\boldsymbol{x}}_{jd} - \bar{\boldsymbol{x}}_{jd} \mid \mathbf{\Delta} \right) \boldsymbol{\beta}_{jd}$$
$$+ \frac{2(n_d - n_{jd})}{n_d} \boldsymbol{\beta}'_{jd} \mathrm{Cov}\left(\bar{y}_{jd}, \tilde{\boldsymbol{x}}_{jd} - \bar{\boldsymbol{x}}_{jd} \mid \mathbf{\Delta} \right)$$
$$= \frac{\sigma_{jd}^2}{n_{dj}} + \frac{(n_d - n_{jd})^2}{n_d^2} \frac{n_d}{(n_d - n_{jd}) n_{jd}} \boldsymbol{\beta}'_{jd} \mathbf{\Sigma}_d \boldsymbol{\beta}_{jd}$$
$$- \frac{2(n_d - n_{jd})}{n_d} \boldsymbol{\beta}'_{jd} \mathrm{Cov}\left(\bar{y}_{jd}, \bar{\boldsymbol{x}}_{jd} \mid \mathbf{\Delta} \right)$$
$$= \frac{\sigma_{jd}^2}{n_{jd}} - \frac{n_d - n_{jd}}{n_d n_{jd}} \boldsymbol{\beta}'_{jd} \mathbf{\Sigma}_d \boldsymbol{\beta}_{jd}.$$

This shows the desired result and completes the proof. ■

Acknowledgments

We would like to thank the two referees for comments. The authors' research was partially supported by the National Natural Science Foundation of China grant 11831008 and the U.S. National Science Foundation grant DMS-1914411.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The authors' research was partially supported by the National Natural Science Foundation of China grant 11831008 and the U.S. National Science Foundation grant DMS-1914411.

Notes on contributors

Dr. Sijing Li holds a Ph.D. in statistics from University of Wisconsin-Madison. She is now a statistician at Roche in Shanghai, China. Her research interest is in missing data.

Dr. Jun Shao holds a PhD in statistics from the University of Wisconsin-Madison. He is a Professor of Statistics at the University of Wisconsin-Madison. His research interests include variable selection and inference with high dimensional data, sample surveys, and missing data problems.

References

Cassel, C. M., Sarndal, C. E., & Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, *63*, 615–620. https://doi.org/10.1093/biomet/63.3.615

Chen, S. X., Leung, D. H., & Qin, J. (2008). Improving semi-parametric estimation by using surrogate data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*, 803–823. https://doi.org/10.1111/rssb.2008.70.issue-4

Greenlees, J. S., Reece, W. S., & Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, *77*, 251–261. https://doi.org/10.1080/01621459.1982.10477793

Hall, A. R. (2005). *Generalized method of moments*. Oxford University Press.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, *50*, 1029–1054. https://doi.org/10.2307/1912775

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley.

Robins, J. M., & Rotiv, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semiparametric models. *Statistics in Medicine*, *16*, 285–319. https://doi.org/10.1002/(ISSN)1097-0258

Rotnitzky, A., & Robins, J. M. (1997). Analysis of semiparametric regression models with nonignorable nonresponse. *Statistics in Medicine*, *16*, 81–102. https://doi.org/10.1002/(ISSN)1097-0258

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592. https://doi.org/10.1093/biomet/63.3.581

Sarndal, C. E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. Springer-Verlag.

Shao, J., & Wang, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, *103*, 175–187. https://doi.org/10.1093/biomet/asv071

Shao, J., & Wang, S. (2014). Efficiency of model-assisted regression estimators in sample surveys. *Statistica Sinica*, *24*, 395–414.

Shao, J., & Zhang, J. (2015). A transformation approach in linear mixed-effect models with informative missing responses. *Biometrika*, *102*, 107–119. https://doi.org/10.1093/biomet/asu069

Tang, G., Little, R. J. A., & Raghunathan, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, *90*, 747–764. https://doi.org/10.1093/biomet/90.4.747

Wang, S., Shao, J., & Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, *24*, 1097–1116.

Wu, M. C., & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, *44*, 175–188. https://doi.org/10.2307/2531905

Xu, L., & Shao, J. (2009). Estimation in longitudinal or panel data models with random-effect-based missing responses. *Biometrics*, *65*, 1175–1183. https://doi.org/10.1111/j.1541-0420.2009.01195.x

Zhao, J., & Shao, J. (2015). Semiparametric pseudo likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, *110*, 1577–1590. https://doi.org/10.1080/01621459.2014.983234