Annals of the Institute of Statistical Mathematics (2023) 75:1–15
https://doi.org/10.1007/s10463-022-00836-4

Estimation with multivariate outcomes having nonignorable item nonresponse

Lyu Ni[1] · **Jun Shao**[1,2]

Received: 26 June 2021 / Revised: 4 February 2022 / Accepted: 31 March 2022 /
Published online: 10 June 2022

Abstract
To estimate unknown population parameters based on $\boldsymbol{y}$, a vector of multivariate outcomes having nonignorable item nonresponse that directly depends on $\boldsymbol{y}$, we propose an innovative inverse propensity weighting approach 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 utilize 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 utilize a modified generalized method of moments to obtain estimators of the parameters in the nonresponse probability. Simulation results are presented and an application is illustrated in a real data set.

Keywords Generalized method of moments · Item nonresponse · Inverse propensity weighting · Multivariate outcome · Nonresponse instrument

1 Introduction

In many statistical applications, multivariate outcomes or responses are collected from every sampled unit in the study. For example, in health studies conducted by the U.S. Centers for Disease Control and Prevention, measurements of total cholesterol, high-density lipoprotein cholesterol, body mass index, average sagittal abdominal diameter, etc. may be obtained from each sampled person in the non-institutionalized civilian resident population of the USA. Longitudinal responses are another type of multivariate outcomes, in which each sampled unit

✉ Jun Shao
shao@stat.wisc.edu

[1] School of Data Science and Engineering, East China Normal University, 3663 North Zhongshan Rd., Shanghai 200050, China

[2] School of Statistics, East China Normal University, 3663 North Zhongshan Road, Shanghai 200050, China

is repeatedly measured over several time periods. An example is the AIDS Clinical Trial Group 193A discussed in Sect. 4 for HIV-AIDS patients with advanced immune suppression.

Unfortunately, item nonresponse is a common phenomenon in multivariate responses, i.e., some of the multivariate responses, not necessarily all, may be missing with a pattern varying with sampled unit. Estimation and statistical inference without taking nonresponse into consideration may lead to seriously biased estimators and conclusions.

Throughout this article, $\boldsymbol{y}$ denotes a k-dimensional outcome or response vector of interest that is subject to item nonresponse, $\boldsymbol{r}$ denotes 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 missing), $j = 1, ..., k$, and $\boldsymbol{x}$ denotes a p-dimensional covariate vector associated with $\boldsymbol{y}$ that is always observed. 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 $p(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x})$. If $p(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x}) = p(\boldsymbol{r}|\boldsymbol{y}_o, \boldsymbol{x})$, where $\boldsymbol{y}_o$ is the observed part of $\boldsymbol{y}$, then nonresponse is ignorable (Rubin 1976; Little and Rubin 2002). Otherwise, nonresponse is nonignorable. While there is a rich literature for valid inference under ignorable nonresponse (Little and Rubin 2002), there are serious challenges under nonignorable nonresponse, especially for multivariate $\boldsymbol{y}$ with item nonresponse.

Greenlees et al. (1982) proposed to handle nonignorable item nonresponse by maximum likelihood estimation, assuming parametric models on both $p(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x})$ and $p(\boldsymbol{y}|\boldsymbol{x})$, the conditional density of $\boldsymbol{y}$ given $\boldsymbol{x}$. However, a fully parametric approach is sensitive to the parametric model assumptions. Since the population $p(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{x}) = p(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})$ is not identifiable when both $p(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x})$ and $p(\boldsymbol{y}|\boldsymbol{x})$ are nonparametric (Robins and Ritov 1997), efforts have been made in scenarios where one of $p(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x})$ and $p(\boldsymbol{y}|\boldsymbol{x})$ is parametric or semi-parametric. Tang et al. (2003) and Zhao and Shao (2015) considered the situation where $p(\boldsymbol{y}|\boldsymbol{x})$ is parametric but $p(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x})$ is nonparametric, whereas Wang et al. (2014) and Shao and Wang (2016) studied a univariate response y ($k = 1$) with a nonparametric $p(y|\boldsymbol{x})$ and a parametric or semi-parametric $p(\boldsymbol{r}|y, \boldsymbol{x})$. Under a mixed-effect model on $p(\boldsymbol{y}|\boldsymbol{x})$, Wu and Carrol (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 conditional density $p(\boldsymbol{y}|\boldsymbol{x})$ and nonparametric marginal density $p(\boldsymbol{y})$, in this paper we propose an innovative inverse propensity weighting approach to construct valid estimators of population parameters in the presence of nonignorable item nonresponse in $\boldsymbol{y}$, assuming the following two assumptions on the propensity:

(A1) The covariate vector $\boldsymbol{x} = (\boldsymbol{u}, z)$ with a non-constant sub-vector z such that $p(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x}) = p(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{u})$ and $p(\boldsymbol{y}|\boldsymbol{x}) = p(\boldsymbol{y}|\boldsymbol{u}, z)$ depends on z.

(A2) Given $(\boldsymbol{y}, \boldsymbol{u})$, components of $\boldsymbol{r}$ are conditionally independent and, for each $j = 1, ..., k$, the probability of observing the jth component of $\boldsymbol{y}$ is $\pi_j(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\theta}_j)$,

where $\boldsymbol{\theta}_j$ is an unknown parameter vector and π_j is a known function of $(\boldsymbol{y}, \boldsymbol{u})$ when $\boldsymbol{\theta}_j$ is known.

The covariate $\boldsymbol{z}$ in (A1) is referred to as a nonresponse instrument (Wang et al. 2014; Zhao and Shao 2015). The existence of a nonresponse instrument that can be excluded from the propensity is almost necessary for handling nonignorable nonresponse (Wang et al. 2014; Zhao and Shao 2015; Shao and Wang 2016). Also, as discussed earlier, the parametric assumption on propensity is needed as $p(\boldsymbol{y}|\boldsymbol{x})$ is nonparametric. Finally, the conditional independence of components of $\boldsymbol{r}$ given $(\boldsymbol{y}, \boldsymbol{u})$ in (A2) is actually reasonable in many applications with item nonresponse, as the conditional independence is not the same as the unconditional independence of components of $\boldsymbol{r}$.

Under (A2), conditioned on $(\boldsymbol{y}, \boldsymbol{u})$, the nonresponse propensity $\pi_j(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\theta}_j)$ not only directly depends on the entire $\boldsymbol{y}$ and possibly $\boldsymbol{u}$, but also varies with j (component). No general result is available under this type of item nonresponse in the literature. The closest is Li and Shao (2022), but it assumes that given $(\boldsymbol{y}, \boldsymbol{u})$, components of $\boldsymbol{r}$ are identically distributed, which may not be realistic when components of $\boldsymbol{y}$ have different distributions (see the real data example in Sect. 4).

Our main methodology is introduced in Sect. 2, followed by some simulation results in Sect. 3 and one real data example in Sect. 4.

2 Methodology

Let $(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{r}_i)$, $i = 1, ..., n$, be identically distributed and independently sampled from the population of $(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{r})$. 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 one. Under assumptions (A1)-(A2), we propose to estimate population parameters using inverse propensity weighting, based on observed data in $(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{r}_i)$, $i = 1, ..., n$.

2.1 Estimation when $\boldsymbol{\theta}_j$'s are known

To illustrate the idea, we consider estimating population mean $\mu_j = E(y_j)$, where y_j is the jth component of $\boldsymbol{y}$ and j is a fixed integer between 1 and k. Estimation of other parameters is discussed in the end of this subsection.

In this subsection, we assume that $\boldsymbol{\theta}_j$'s in (A2) are known. Estimation of $\boldsymbol{\theta}_j$'s is considered in the next subsection. For $\boldsymbol{r}_i$ and $\boldsymbol{y}_i$, denote their jth components by r_{ij} and y_{ij}, respectively. The simple inverse propensity weighting estimator,

$$\sum_{i=1}^{n} \frac{r_{ij} y_{ij}}{\pi_j(\boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{\theta}_j)} \Big/ \sum_{i=1}^{n} \frac{r_{ij}}{\pi_j(\boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{\theta}_j)},$$

which works for the univariate case of $k = 1$, does not work because $\pi_j(\boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{\theta}_j)$ cannot be computed when $\boldsymbol{y}_i$ has a missing component $l \neq j$. Thus, we propose the following estimator of μ_j using composite inverse propensity weighting:

$$\widehat{\mu}_j(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{(r_{i1}\cdots r_{ik})\, y_{ij}}{\pi_1(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta}_1)\cdots\pi_k(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta}_k)} \Big/ \sum_{i=1}^{n} \frac{r_{i1}\cdots r_{ik}}{\pi_1(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta}_1)\cdots\pi_k(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta}_k)}, \quad (1)$$

where $\boldsymbol{\theta}$ is a vector with $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k$ as sub-vectors assumed to be known at this moment. Since the product $r_{i1}\cdots r_{ik}$ is used, we must use the product $\pi_1(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta}_1)\cdots\pi_k(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta}_k)$ as weights, which can be computed when $r_{i1}\cdots r_{ik}=1$. To see why $\widehat{\mu}_j(\boldsymbol{\theta})$ in (1) is asymptotically valid as $n\to\infty$, note that

$$\begin{aligned}
&E\left\{\frac{(r_{i1}\cdots r_{ik})\, y_{ij}}{\pi_1(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta}_1)\cdots\pi_k(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta}_k)}\right\}\\
&= E\left[E\left\{\frac{(r_{i1}\cdots r_{ik})\, y_{ij}}{\pi_1(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta}_1)\cdots\pi_k(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta}_k)}\Big|\boldsymbol{y}_i,\boldsymbol{u}_i\right\}\right]\\
&= E\left[\frac{y_{ij}E(r_{i1}\cdots r_{ik}|\boldsymbol{y}_i,\boldsymbol{u}_i)}{\pi_1(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta}_1)\cdots\pi_k(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta}_k)}\right]\\
&= E\left[\frac{y_{ij}E(r_{i1}|\boldsymbol{y}_i,\boldsymbol{u}_i)\cdots E(r_{ik}|\boldsymbol{y}_i,\boldsymbol{u}_i)}{\pi_1(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta}_1)\cdots\pi_k(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta}_k)}\right]\\
&= E(y_{ij}) = \mu_j,
\end{aligned}$$

where the third equality follows from the independence of r_{ij}'s conditioned on $(\boldsymbol{y}_i,\boldsymbol{u}_i)$ and the last equality follows from $E(r_{ij}|\boldsymbol{y}_i,\boldsymbol{u}_i)=\pi_j(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta}_j)$, under (A1)–(A2). The consistency and asymptotic normality of $\widehat{\mu}_j(\boldsymbol{\theta})$ as $n\to\infty$ can be established by applying standard arguments and the central limit theorem, under some moment conditions, since the right hand side of (1) is a ratio of sums of independent random variables.

In this way, other population characteristics can be similarly estimated. For example, if we want to estimate the distribution of the jth 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 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 defined by some estimating equations.

2.2 Estimation of $\boldsymbol{\theta}$

To complete our proposed methodology, we need to remove the assumption that $\boldsymbol{\theta}$ is known, by constructing an estimator $\widehat{\boldsymbol{\theta}}_j$ of $\boldsymbol{\theta}_j$ for each j under (A1)–(A2). To estimate $\boldsymbol{\theta}_j$, we follow the approach of generalized method of moments (GMM) in Wang et al. (2014) for the univariate response, but we need to add a novel modification to handle the multivariate $\boldsymbol{y}$.

A brief description of the GMM is as follows. Let $\boldsymbol{\varphi}$ be the parameter vector to estimate, which is a unique solution to $E\{\boldsymbol{g}(\boldsymbol{\varphi})\}=0$ with an l-dimensional vector estimating function $\boldsymbol{g}$ whose tth component is $g_t(\boldsymbol{y},\boldsymbol{x},\boldsymbol{r},\boldsymbol{\varphi})$, $t=1,...,l$. The functions $g_1,...,g_l$ are chosen so that l is not less than the dimension of $\boldsymbol{\varphi}$ and at the true parameter value $\boldsymbol{\varphi}$, $E\{\partial\boldsymbol{g}(\boldsymbol{\varphi})/\partial\boldsymbol{\varphi}\}$ is of full rank. Let $\boldsymbol{g}_n(\boldsymbol{\varphi})$ be the l-dimensional vector

whose tth component is the sample average $n^{-1}\sum_{i=1}^{n} g_t(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{r}_i, \boldsymbol{\varphi})$, $t = 1, ..., l$. If l is the same as the dimension of $\boldsymbol{\varphi}$, then we estimate $\boldsymbol{\varphi}$ by $\widehat{\boldsymbol{\varphi}}$ such that $\boldsymbol{g}_n(\widehat{\boldsymbol{\varphi}}) = 0$. If l is larger than the dimension of $\boldsymbol{\varphi}$, we apply the following two-step GMM (Hansen 1982; Hall 2005):

1. Obtain $\widetilde{\boldsymbol{\varphi}}$ by minimizing $\{\boldsymbol{g}_n(\boldsymbol{\varphi})\}^T \boldsymbol{g}_n(\boldsymbol{\varphi})$, where $\boldsymbol{a}^T$ is the transpose of column vector $\boldsymbol{a}$.
2. Obtain $\widehat{\boldsymbol{\varphi}}$ by minimizing $\{\boldsymbol{g}_n(\boldsymbol{\varphi})\}^T \widehat{\boldsymbol{W}} \boldsymbol{g}_n(\boldsymbol{\varphi})$, where $\widehat{\boldsymbol{W}}$ is the inverse of $l \times l$ matrix whose (t, t') element is $n^{-1}\sum_{i=1}^{n} g_t(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{r}_i, \widetilde{\boldsymbol{\varphi}}) g_{t'}(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{r}_i, \widetilde{\boldsymbol{\varphi}})$.

The optimization can be solved by using the MATLAB or R function fminsearch.

For our problem, it remains to specify the form of the estimating function $\boldsymbol{g}$. To fix the idea, suppose first that the nonresponse instrument $\boldsymbol{z}$ is discrete and has s categories, say $\boldsymbol{z} \in \{z_1, ..., z_s\}$. A straightforward extension of the approach in Wang et al. (2014) (from univariate response to multivariate $\boldsymbol{y}$) is using

$$\boldsymbol{g}(\boldsymbol{\theta}) = \left\{ \frac{r_1 \cdots r_k}{\pi_1(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\theta}_1) \cdots \pi_k(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\theta}_k)} - 1 \right\} \boldsymbol{v}, \tag{2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, ..., \boldsymbol{\theta}_k^T)^T$, r_j is the jth component of the vector $\boldsymbol{r}$ of response indicators and $\boldsymbol{v}$ is the $(s + q)$-dimensional vector whose first s components are indicators of $\boldsymbol{z} = z_t$, $t = 1, ..., s$, and the rest q components are the q-dimensional covariate vector $\boldsymbol{u}$ in (A1)–(A2). With this choice of $\boldsymbol{g}$, $E\{\boldsymbol{g}(\boldsymbol{\theta})\} = 0$ under (A1)–(A2).

However, there is a problem: $l = s + q$ may be smaller than $\dim(\boldsymbol{\theta})$, the dimension of $\boldsymbol{\theta}$. For example, if $\boldsymbol{u}$ is continuous and

$$\pi_j(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\theta}_j) = \{1 + \exp(\alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{y} + \boldsymbol{\gamma}_j^T \boldsymbol{u})\}^{-1}, \qquad j = 1, ..., k, \tag{3}$$

where α_j is univariate, $\boldsymbol{\beta}_j$ is k-dimensional, $\boldsymbol{\gamma}_j$ is q-dimensional, and $\boldsymbol{\theta}_j = (\alpha_j, \boldsymbol{\beta}_j^T, \boldsymbol{\gamma}_j^T)^T$ with $\dim(\boldsymbol{\theta}_j) = q + k + 1$, then $l = s + q \geq k(q + k + 1) = \dim(\boldsymbol{\theta})$ means that $s \geq (k - 1)q + k(k + 1)$, which may be unrealistic. For instance, when $q = 0$ (there is no $\boldsymbol{u}$), $s \geq k(k + 1)$ requires that $\boldsymbol{z}$ has at least $k(k + 1)$ categories.

To overcome this difficulty we consider the following modification. First, we construct k overlapped sub-sets $D_1, ..., D_k$ of the entire data set, where D_h contains data from units whose y_{ih} may be missing but all other components are observed, $h = 1, ..., k$. 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 in the case of $k = 3$ and $n = 30$.

Then, we estimate $\boldsymbol{\theta}_j$ one at a time, $j = 1, ..., k$. For each j, we use data in D_j and estimating function

$$\boldsymbol{g}^{(j)}(\boldsymbol{\theta}_j) = \left\{ \frac{r_j}{\pi_j(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\theta}_j)} - 1 \right\} \delta_j \boldsymbol{v}_j, \tag{4}$$

where δ_j is the indicator of set D_j, $\boldsymbol{v}_j$ is the vector whose first $s + q$ components are the same as those of $\boldsymbol{v}$ in (2), the rest $k - 1$ components are $y_1, ..., y_{j-1}, y_{j+1}, ..., y_k$, and

Table 1 Example of D_1, D_2, D_3 when $k = 3$ and $n = 30$ (r_j is the indicator of whether y_j is observed)

Entire data set				D_1				D_2				D_3			
unit	r_1	r_2	r_3	unit	r_1	r_2	r_3	unit	r_1	r_2	r_3	unit	r_1	r_2	r_3
1	0	0	0	2	1	1	1	2	1	1	1	2	1	1	1
2	1	1	1	3	0	1	1	5	1	1	1	5	1	1	1
3	0	1	1	5	1	1	1	8	1	1	1	7	1	1	0
4	1	0	0	8	1	1	1	11	1	0	1	8	1	1	1
5	1	1	1	12	1	1	1	12	1	1	1	10	1	1	0
6	0	0	1	15	1	1	1	14	1	0	1	12	1	1	1
7	1	1	0	16	0	1	1	15	1	1	1	15	1	1	1
8	1	1	1	17	1	1	1	17	1	1	1	17	1	1	1
9	0	0	1	21	1	1	1	21	1	1	1	18	1	1	0
10	1	1	0	23	1	1	1	22	1	0	1	20	1	1	0
11	1	0	1	24	0	1	1	23	1	1	1	21	1	1	1
12	1	1	1	28	1	1	1	27	1	0	1	23	1	1	1
13	0	1	0					28	1	1	1	28	1	1	1
14	1	0	1					30	1	0	1	29	1	1	0
15	1	1	1												
16	0	1	1												
17	1	1	1												
18	1	1	0												
19	0	0	0												
20	1	1	0												
21	1	1	1												
22	1	0	1												
23	1	1	1												
24	0	1	1												
25	0	0	0												
26	0	1	0												
27	1	0	1												
28	1	1	1												
29	1	1	0												
30	1	0	1												

y_t is the tth component of $\boldsymbol{y}$. A GMM estimator $\widehat{\boldsymbol{\theta}}_j$ of $\boldsymbol{\theta}_j$ can be computed using the estimating function $\boldsymbol{g}^{(j)}$ in (4) and data set D_j. Note that $\boldsymbol{g}^{(j)}(\boldsymbol{\theta}_j)$ in (4) can always be computed, since when $\delta_j = 1$, all y_t with $t \neq j$ are observed.

To see why the function $\boldsymbol{g}^{(j)}(\boldsymbol{\theta}_j)$ in (4) produces asymptotically valid estimator of $\boldsymbol{\theta}_j$, note that

$$\begin{aligned} E\{\boldsymbol{g}^{(j)}(\boldsymbol{\theta}_j)\} &= E\left[E\left[\left\{\frac{r_j}{\pi_j(\boldsymbol{y},\boldsymbol{u},\boldsymbol{\theta}_j)}-1\right\}\delta_j\boldsymbol{v}_j\middle|\boldsymbol{y},\boldsymbol{u},\delta_j\right]\right] \\ &= E\left[\left\{\frac{E(r_j|\boldsymbol{y},\boldsymbol{u},\delta_j)}{\pi_j(\boldsymbol{y},\boldsymbol{u},\boldsymbol{\theta}_j)}-1\right\}\delta_j E(\boldsymbol{v}_j|\boldsymbol{y},\boldsymbol{u},\delta_j)\right] \\ &= 0, \end{aligned}$$

where the second equality follows from the independence between $\boldsymbol{z}$ and r_j conditioned on $(\boldsymbol{y},\boldsymbol{u},\delta_j)$ and the last equality follows from $E(r_j|\boldsymbol{y},\boldsymbol{u},\delta_j) = E(r_j|\boldsymbol{y},\boldsymbol{u}) = \pi_j(\boldsymbol{y},\boldsymbol{u},\boldsymbol{\theta}_j)$ under (A1)–(A2).

A key difference between $\boldsymbol{g}$ and $\boldsymbol{g}^{(j)}$ is that the observed components of $\boldsymbol{y}$ other than the jth component are used as "covariates" and included in the vector $\boldsymbol{v}_j$ in (4). In this way, we not only make use of the partially observed responses in $\boldsymbol{y}$ (note that $r_1\cdots r_k = 1$ if and only if all components of $\boldsymbol{y}$ are observed), but also include more components in the estimating function so that l is typically large enough for our purpose of estimating $\boldsymbol{\theta}_j$. For example, in the case of (3), $\dim(\boldsymbol{\theta}_j) = q+k+1$; hence, $l = s+q+k-1 \geq q+k+1$ is the same as $s \geq 2$, which naturally holds as long as $\boldsymbol{z}$ is not a constant. However, if we do not include the last $k-1$ components in $\boldsymbol{v}_j$, i.e., $\boldsymbol{v}_j$ in (4) is replaced by $\boldsymbol{v}$ defined in (2), then the dimension of $\boldsymbol{g}^{(j)}$ is $s+q$, which is smaller than the dimension of $\boldsymbol{\theta}_j$ in the case of (3) unless $s \geq k+1$. Therefore, using $\boldsymbol{v}_j$ instead of $\boldsymbol{v}$ ensures that our procedure has a larger scope in application.

Note that the estimating function $\boldsymbol{g}$ in (2) involves $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T,...,\boldsymbol{\theta}_k^T)^T$ and the estimating function $\boldsymbol{g}^{(j)}$ in (4) involves $\boldsymbol{\theta}_j$ only, i.e., we decompose the estimation of a parameter vector with dimension $\dim(\boldsymbol{\theta}) = \sum_{j=1}^k \dim(\boldsymbol{\theta}_j)$ into k estimation problems, each with dimension $\dim(\boldsymbol{\theta}_j)$. Even if $l \geq \dim(\boldsymbol{\theta})$ and simultaneous estimation of $\boldsymbol{\theta}$ is possible, the large dimension of $\boldsymbol{\theta}$ in GMM may result in numerical unstableness or inaccuracy. Furthermore, it is clear that each D_j contains the set with $r_1\cdots r_k = 1$ used in (2) and, thus, estimating $\boldsymbol{\theta}_j$'s separately utilizes more data, although some data are repeatedly used since D_j's are overlapped.

When $\boldsymbol{z}$ is continuous, we can define $\boldsymbol{v}$ in (2) to be the vector of first s moments of $\boldsymbol{z}$. For example, if $\boldsymbol{z} = z$ is univariate, then we use $\boldsymbol{v} = (1,z)^T$ with $s = 2$; if $\boldsymbol{z}$ is bivariate with components z_1 and z_2, then $\boldsymbol{v} = (1,z_1,z_2)^T$ with $s = 3$. We can also apply the method by discretizing $\boldsymbol{z}$ into s categories with approximately equal sizes and a small s.

Once $\widehat{\boldsymbol{\theta}}_1,...,\widehat{\boldsymbol{\theta}}_k$ are obtained, we estimate μ_j by $\widehat{\mu}_j(\widehat{\boldsymbol{\theta}})$, obtained by substituting $\boldsymbol{\theta}$ in $\widehat{\mu}_j(\boldsymbol{\theta})$ in (1) with $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\theta}}_1^T,...,\widehat{\boldsymbol{\theta}}_k^T)^T$.

2.3 Asymptotic theory

Under the same regularity conditions assumed in Wang et al. (2014), consistency and asymptotic normality of $\widehat{\boldsymbol{\theta}}_j$ can be established and details are omitted. For the point estimator $\widehat{\mu}_j(\widehat{\boldsymbol{\theta}})$, its consistency and asymptotic normality can be established. We provide the main argument below and omit the details of proof. Define

$$\psi(\boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{\theta}) = \frac{1}{\pi_1(\boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{\theta}_1) \cdots \pi_k(\boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{\theta}_k)},$$
$$\tau(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (r_{i1} \cdots r_{ik})\, \psi(\boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{\theta}),$$

and

$$\zeta_j(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (r_{i1} \cdots r_{ik}) y_{ij}\, \psi(\boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{\theta}).$$

Then, by (1), $\widehat{\mu}_j(\widehat{\boldsymbol{\theta}}) = \zeta_j(\widehat{\boldsymbol{\theta}})/\tau(\widehat{\boldsymbol{\theta}})$ and

$$\sqrt{n}\{\widehat{\mu}_j(\widehat{\boldsymbol{\theta}}) - \mu_j\} = \frac{1}{\tau(\widehat{\boldsymbol{\theta}})} \Big[\sqrt{n}\{\zeta_j(\widehat{\boldsymbol{\theta}}) - \mu_j\} - \mu_j \sqrt{n}\{\tau(\widehat{\boldsymbol{\theta}}) - 1\} \Big].$$

Assume that $\nabla\psi(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\theta}) = \partial\psi(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}$ exists and each component of $\nabla\psi(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\vartheta}) - \nabla\psi(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\theta})$ is bounded in absolute value by $H(\boldsymbol{y}, \boldsymbol{u})\|\boldsymbol{\vartheta} - \boldsymbol{\theta}\|$ with $E\{H(\boldsymbol{y}, \boldsymbol{u})\} < \infty$, where $\|\cdot\|$ is the L_2 norm. This assumption holds if $\pi_j(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\theta}_j)$'s are given by (3). Then, by the consistency of the GMM estimator $\widehat{\boldsymbol{\theta}}$,

$$\begin{aligned}
\sqrt{n}\{\zeta_j(\widehat{\boldsymbol{\theta}}) - \mu_j\} &= \sqrt{n}\{\zeta_j(\boldsymbol{\theta}) - \mu_j\} + \sqrt{n}\{\zeta_j(\widehat{\boldsymbol{\theta}}) - \zeta_j(\boldsymbol{\theta})\} \\
&= \sqrt{n}\{\zeta_j(\boldsymbol{\theta}) - \mu_j\} + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (r_{i1} \cdots r_{ik}) y_{ij} \Big\{ \psi(\boldsymbol{y}_i, \boldsymbol{u}_i, \widehat{\boldsymbol{\theta}}) - \psi(\boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{\theta}) \Big\} \\
&= \sqrt{n}\{\zeta_j(\boldsymbol{\theta}) - \mu_j\} + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (r_{i1} \cdots r_{ik}) y_{ij} \nabla\psi(\boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o_p(1) \\
&= \sqrt{n}\{\zeta_j(\boldsymbol{\theta}) - \mu_j\} + \left\{ \frac{1}{n} \sum_{i=1}^{n} (r_{i1} \cdots r_{ik}) y_{ij} \nabla\psi(\boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{\theta}) \right\} \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o_p(1) \\
&= \sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^{n} (r_{i1} \cdots r_{ik}) y_{ij}\, \psi(\boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{\theta}) - \mu_j \right\} + A(\boldsymbol{\theta}) \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o_p(1),
\end{aligned}$$

where $A(\boldsymbol{\theta}) = E\{(r_{i1} \cdots r_{ik}) y_{ij} \nabla\psi(\boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{\theta})\}$, $o_p(1)$ denotes a term converging to 0 in probability as $n \to \infty$, and the last equality follows from the law of large numbers and the definition of $\zeta_j(\boldsymbol{\theta})$. Similarly,

$$\begin{aligned}\sqrt{n}\{\tau(\widehat{\boldsymbol{\theta}})-1\} &= \sqrt{n}\{\tau(\boldsymbol{\theta})-1\}+\sqrt{n}\{\tau(\widehat{\boldsymbol{\theta}})-\tau(\boldsymbol{\theta})\}\\ &= \sqrt{n}\{\tau(\boldsymbol{\theta})-1\}+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(r_{i1}\cdots r_{ik})\Big\{\psi(\boldsymbol{y}_i,\boldsymbol{u}_i,\widehat{\boldsymbol{\theta}})-\psi(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta})\Big\}\\ &= \sqrt{n}\{\tau(\boldsymbol{\theta})-1\}+\left\{\frac{1}{n}\sum_{i=1}^{n}(r_{i1}\cdots r_{ik})\nabla\psi(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta})\right\}\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})+o_p(1)\\ &= \sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}(r_{i1}\cdots r_{ik})\psi(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta})-1\right\}+B(\boldsymbol{\theta})\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})+o_p(1),\end{aligned}$$

where $B(\boldsymbol{\theta})=E\{(r_{i1}\cdots r_{ik})\nabla\psi(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta})\}$. From the theory in Wang et al. (2014), the GMM estimator $\widehat{\boldsymbol{\theta}}$ has the property that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\phi}(\boldsymbol{y}_i,\boldsymbol{x}_i)+o_p(1), \tag{5}$$

where $\boldsymbol{\phi}$ is an unknown vector function with $E\{\boldsymbol{\phi}(\boldsymbol{y},\boldsymbol{x})\}=0$ and a finite positive definite matrix $E\{\boldsymbol{\phi}(\boldsymbol{y},\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{y},\boldsymbol{x})^T\}$. Then, the asymptotic normality of $\sqrt{n}\{\widehat{\mu}_j(\widehat{\boldsymbol{\theta}})-\mu_j\}$ with asymptotic mean 0 follows from the joint asymptotic normality of the following vector,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}(r_{i1}\cdots r_{ik})y_{ij}\,\psi(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta})-\mu_j\\ (r_{i1}\cdots r_{ik})\,\psi(\boldsymbol{y}_i,\boldsymbol{u}_i,\boldsymbol{\theta})-1\\ \boldsymbol{\phi}(\boldsymbol{y}_i,\boldsymbol{x}_i)\end{pmatrix}. \tag{6}$$

However, the asymptotic variance of $\sqrt{n}\{\widehat{\mu}_j(\widehat{\boldsymbol{\theta}})-\mu_j\}$ is very complicated, because it involves not only the asymptotic variances of the three components in (6), but also their asymptotic covariances, and the form of function $\boldsymbol{\phi}$ in (5) is complicated (Wang et al. 2014). Thus, we do not try to obtain an explicit form of the asymptotic variance of $\widehat{\mu}_j(\widehat{\boldsymbol{\theta}})$. Instead, we recommend the bootstrap method for variance estimation or inference. Using the previous arguments, we can show that the bootstrap analog $\widehat{\mu}_j^*(\widehat{\boldsymbol{\theta}}^*)$ is asymptotically normal and the general bootstrap theory (Shao and Tu 1995) ensures that the bootstrap variance estimators are consistent. Applying the bootstrap effectively avoids the complicated derivation of asymptotic variances, at the expense of a large amount of computations. In Sect. 3, the performance of bootstrap standard error (squared root of variance estimator) is evaluated by simulations.

2.4 Discussion

Although our proposed estimators are asymptotically valid, the form of $\boldsymbol{v}$ in (2) is not unique. An interesting but difficult research problem is whether there exists a choice of $\boldsymbol{v}$ optimal in some sense. Another discussion is whether there are sub-sets of data other than $D_1,...,D_k$ for the purpose of estimating $\boldsymbol{\theta}_j$'s. There are two key issues. First, the estimating function $\boldsymbol{g}^{(j)}(\boldsymbol{\theta}_j)$ in (4) needs to be computed based on

the sub-sets, which is not simple because $\pi_j(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\theta}_j)$ in (4) depends on the entire $\boldsymbol{y}$ whose components may be missing. Second, the estimating function should satisfy $E\{\boldsymbol{g}^{(j)}(\boldsymbol{\theta}_j)\} = 0$, as we show in Sect. 2.2 for our suggested D_j. This makes the choice of sub-sets very limited due to the nonignorable item nonresponse.

3 Simulation results

We carry out a simulation study under three settings to investigate the finite sample performance of our proposed $\widehat{\mu}_j(\widehat{\boldsymbol{\theta}})$ given by (1) as an estimator of the marginal population mean $\mu_j = E(y_j)$, $j = 1, ..., k$, with $\boldsymbol{\theta}$ estimated by the GMM estimator $\widehat{\boldsymbol{\theta}}$ derived in Sect. 2.2.

In the first two settings, we consider a panel size $k = 4$ and sample size $n = 1,200$, reflexing the panel and sample sizes in the real data AIDS Clinical Trial Group 193A example presented in Sect. 4. A univariate and continuous covariate z is considered with $\log z \sim N(2.9, 1)$. Given z, y_j's are conditionally independent, $\log y_1 \sim N(0.4 + 0.9 \log z, 0.8^2)$, $\log y_2 \sim N(0.6 + 0.8 \log z, 0.8^2)$, $\log y_3 \sim N(0.8 + 0.7 \log z, 0.8^2)$, and $\log y_4 \sim N(0.9 + 0.6 \log z, 0.8^2)$. The true marginal means are $\mu_1 = 41.89$, $\mu_2 = 35.16$, $\mu_3 = 29.81$, and $\mu_4 = 23.10$. These μ_j's are chosen to match the estimated values in the real data example considered in Sect. 4.

Table 2 Simulation results for the estimation of μ_j in setting 1 ($n = 1,200$, bootstrap size = 200, simulation runs = 1,000)

j	% of Missing	Method	Estimate	Bias	Bias %	Standard deviation	Standard error	Coverage prob.
1	30.75	Proposed	42.44	0.552	0.013	7.705	7.896	0.932
		Naive	39.76	−2.129	-0.051	2.448	2.441	0.784
		Full data	41.92	0.034	0.001	2.180	2.150	0.936
2	24.57	Proposed	35.76	0.599	0.017	6.246	6.223	0.934
		Naive	33.92	−1.241	−0.035	1.836	1.792	0.827
		Full data	35.21	0.048	0.001	1.636	1.628	0.943
3	46.66	Proposed	30.75	0.933	0.031	6.035	5.911	0.956
		Naive	28.09	−1.720	−0.058	1.582	1.583	0.747
		Full data	29.91	0.095	0.003	1.236	1.245	0.948
4	39.03	Proposed	23.65	0.543	0.024	4.355	4.317	0.956
		Naive	22.06	−1.040	−0.045	1.079	1.053	0.780
		Full data	23.06	−0.048	−0.002	0.908	0.861	0.932

$\pi_j(\boldsymbol{y}, \boldsymbol{\theta}_j) = \{1 + \exp(\alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{y})\}^{-1}$, $j = 1, ..., 4$
$\alpha_1 = -1.2$, $\boldsymbol{\beta}_1^T = (0.1, 0.01, 0.01, 0.01)$
$\alpha_2 = -1.5$, $\boldsymbol{\beta}_2^T = (0.01, 0.1, 0.01, 0.01)$
$\alpha_3 = -0.5$, $\boldsymbol{\beta}_3^T = (0.01, 0.01, 0.1, 0.01)$
$\alpha_4 = -0.8$, $\boldsymbol{\beta}_4^T = (0.01, 0.01, 0.01, 0.1)$

Table 3 Simulation results for the estimation of μ_j in setting 2 ($n = 1,200$, bootstrap size = 200, simulation runs = 1,000)

j	% of Missing	Method	Estimate	Bias	Bias %	Standard deviation	Standard error	Coverage prob.
1	30.47	Proposed	42.26	0.375	0.009	6.979	7.176	0.941
		Naive	39.58	−2.310	−0.055	2.464	2.382	0.769
		Full data	41.98	0.090	0.002	2.138	2.145	0.943
2	22.79	Proposed	35.63	0.471	0.013	4.891	5.498	0.961
		Naive	35.96	0.796	0.023	1.916	1.857	0.946
		Full data	35.18	0.017	0.001	1.682	1.604	0.929
3	48.19	Proposed	30.14	0.324	0.011	4.699	5.292	0.960
		Naive	28.40	−1.413	−0.047	1.656	1.598	0.797
		Full data	29.87	0.052	0.002	1.264	1.227	0.935
4	40.08	Proposed	23.53	0.426	0.018	3.734	3.989	0.951
		Naive	22.96	0.854	0.037	1.154	1.150	0.926
		Full data	23.12	0.019	0.001	0.879	0.862	0.946

$\pi_j(\boldsymbol{y}, \boldsymbol{\theta}_j) = \{1 + \exp(\alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{y})\}^{-1}$, $j = 1, ..., 4$
$\alpha_1 = -1.3$, $\boldsymbol{\beta}_1^T = (0.1, 0.02, 0.02, 0.02)$
$\alpha_2 = -1.1$, $\boldsymbol{\beta}_2^T = (0.02, -0.1, 0.02, 0.02)$
$\alpha_3 = -0.3$, $\boldsymbol{\beta}_3^T = (-0.02, 0.02, 0.1, -0.02)$
$\alpha_4 = -0.2$, $\boldsymbol{\beta}_4^T = (0.02, 0.02, -0.02, -0.1)$

The nonresponse propensity is given by (3) with $\boldsymbol{u} = 0$, α_j's and $\boldsymbol{\beta}_j$'s shown in Tables 2 and 3 for settings 1–2, respectively. The parameter values α_j and $\boldsymbol{\beta}_j$ are chosen so that the unconditional nonresponse probability matches the observed proportion in the real data example for every j. The difference between two settings is that all coefficients in front of y_j's in the propensity (3) are positive in setting 1 so that larger values of y_j have a higher probability to be nonresponse, whereas in setting 2, the coefficients may be positive or negative. The covariate z in the real data example is the baseline response and is used as nonresponse instrument in the estimation.

In setting 3, we consider a discrete instrument z with three categories, $P(z = 1) = 0.4$, $P(z = 2) = 0.3$, and $P(z = 3) = 0.3$, an additional continuous covariate $u \sim N(2, 1)$, and larger panel and sample sizes, $k = 6$ and $n = 2,000$. Given z and u, y_j's are conditionally independent, $y_1 \sim N(1 + z + u, 1)$, $y_2 \sim N(z + 2u, 1)$, $y_3 \sim N(1 + 2z + u, 1)$, $y_4 \sim N(1 + 2z + 2u, 1)$, $y_5 \sim N(3 + 3z + u, 1)$, and $y_6 \sim N(3 + 3z + 2u, 1)$. The true marginal means are $\mu_1 = 4.9$, $\mu_2 = 5.9$, $\mu_3 = 6.8$, $\mu_4 = 8.8$ and $\mu_5 = 10.7$, and $\mu_6 = 12.7$. The nonresponse propensity is given by (3) with α_j, $\boldsymbol{\beta}_j$, and γ_j specified in Table 4.

To evaluate the performance, we include two other estimators, the naive estimator = the sample mean of observed values of y_j and the sample mean of y_j with full data (no nonresponse) available in the simulation as nonresponse is constructed. The naive estimator is theoretically biased due to nonignorable nonresponse and is included to see the effect of bias; the full data sample mean is used as a standard.

Table 4 Simulation results for the estimation of μ_j in setting 3 ($n = 2,000$, bootstrap size = 200, simulation runs = 1,000)

j	% of Missing	Method	Estimate	Bias	Bias %	Standard deviation	Standard error	Coverage prob.
1	16.87	Proposed	4.857	−0.043	−0.009	0.132	0.148	0.959
		Naive	4.817	−0.083	−0.017	0.039	0.040	0.453
		Full data	4.898	−0.002	−0.000	0.036	0.037	0.946
2	12.44	Proposed	5.887	−0.013	−0.002	0.133	0.147	0.955
		Naive	5.808	−0.092	−0.016	0.057	0.056	0.604
		Full data	5.899	−0.000	−0.000	0.054	0.053	0.950
3	11.49	Proposed	6.803	0.003	0.000	0.129	0.143	0.969
		Naive	6.740	−0.060	−0.009	0.051	0.052	0.795
		Full data	6.798	−0.002	−0.000	0.049	0.049	0.945
4	16.21	Proposed	8.780	−0.020	−0.002	0.184	0.196	0.963
		Naive	8.647	−0.153	−0.017	0.069	0.067	0.360
		Full data	8.798	−0.002	−0.000	0.063	0.062	0.942
5	17.36	Proposed	10.68	−0.024	−0.002	0.168	0.189	0.963
		Naive	10.54	−0.161	−0.015	0.070	0.070	0.359
		Full data	10.80	−0.003	−0.000	0.064	0.064	0.952
6	24.06	Proposed	12.60	−0.101	−0.008	0.218	0.243	0.946
		Naive	12.37	−0.329	−0.026	0.085	0.084	0.031
		Full data	12.70	−0.002	−0.000	0.076	0.075	0.939

$\pi_j(\boldsymbol{y}, \boldsymbol{\theta}_j) = \{1 + \exp(\alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{y} + \gamma_j u)\}^{-1}$, $j = 1, ..., 6$
$\alpha_1 = -2.8$, $\boldsymbol{\beta}_1^T = (0.1, -0.02, 0.02, 0.02, 0.02, 0.02)$, $\gamma_1 = 0.01$
$\alpha_2 = -2.8$, $\boldsymbol{\beta}_2^T = (0.02, 0.1, -0.02, 0.02, -0.02, 0.02)$, $\gamma_2 = 0.02$
$\alpha_3 = -2.8$, $\boldsymbol{\beta}_3^T = (0.02, 0.02, 0.1, -0.02, 0.02, -0.02)$, $\gamma_3 = 0.03$
$\alpha_4 = -2.8$, $\boldsymbol{\beta}_4^T = (0.02, -0.02, 0.02, 0.1, -0.02, 0.02)$, $\gamma_4 = 0.04$
$\alpha_5 = -2.8$, $\boldsymbol{\beta}_5^T = (0.02, -0.02, 0.02, 0.02, 0.1, -0.02)$, $\gamma_5 = 0.05$
$\alpha_6 = -2.8$, $\boldsymbol{\beta}_6^T = (0.02, -0.02, -0.02, 0.02, 0.02, 0.1)$, $\gamma_6 = 0.05$

Based on 1000 simulation runs, Tables 2, 3 and 4 report, for settings 1–3, respectively, simulation average of estimates of μ_j, bias, bias in percentage, standard deviation of the estimate, average of the standard error obtained by bootstrapping, and coverage probability of the approximate 95% confidence interval with limits = estimate $\pm$ 1.96(bootstrap standard error). Results are given for $j = 1, ..., k$ and three estimators, based on the proposed, naive, and full data methods. In the calculation of the proposed estimator given by (1), the GMM estimator $\widehat{\boldsymbol{\theta}}_j$ is calculated using the MATLAB or R function `fminsearch` with initial value $\boldsymbol{\theta}_j = 0$. In settings 1–2, z is continuous and we use $(1, z)^T$ as the first two components of $\boldsymbol{v}_j$ in (4). In setting 3, z is discrete and we use the indicators of three categories of z as the first three components of $\boldsymbol{v}_j$.

From the simulation results in Tables 2, 3 and 4, the performance of proposed estimator (1) can be summarized as follows. It has negligible bias: the largest biases

are 3.1% and 2.4% and the rest of biases are all smaller than 2%, in absolute value. The coverage probability of the related confidence interval is close to 95%; the worst cases are in setting 1: 0.932 when $j = 1$ and 0.934 when $j = 2$, but even the full data approach may also have coverage probabilities 0.932 and 0.936. The bootstrap standard error for the proposed method performs well in general, and is sometimes a little bit conservative, which results in slightly conservative coverage probability of the confidence interval.

In setting 1 where larger y_j values have higher probability to be missing data, the naive estimator has a negative bias. Although the bias is often small, it still affects considerably the coverage probability of the related confidence interval. In setting 2, when smaller y_j values have higher probability to be missing data ($j = 2$ or 4), the naive estimator has a small positive bias = 2.3% and 3.7% so that its coverage probability is acceptable. This appears by luck but cannot support the naive approach of ignoring nonignorable nonresponse.

4 A real data example

For illustration, we apply our proposed estimation method to the AIDS Clinical Trial Group 193A data set, which can be found at https://www.hsph.harvard.edu/fitzmaur/ala/cd4.txt. Longitudinal responses, the CD4 cell counts, were collected from HIV-AIDS patients with advanced immune suppression. After removing some

Table 5 Nonresponse pattern of $\boldsymbol{y}$ in the example

Nonresponse pattern					Number of
r_1	r_2	r_3	r_4	$\sum_j r_j$	observations
0	0	0	0	0	121
1	0	0	0	1	82
0	1	0	0	1	57
0	0	1	0	1	11
0	0	0	1	1	7
1	1	0	0	2	69
1	0	1	0	2	37
1	0	0	1	2	11
0	1	1	0	2	41
0	1	0	1	2	119
0	0	1	1	2	2
1	1	1	0	3	94
1	1	0	1	3	118
1	0	1	1	3	34
0	1	1	1	3	31
1	1	1	1	4	437
Total					1271

Table 6 Estimates and standard errors of μ_j's in the example

j	% of Missing	Method	Estimate	Standard error
1	30.61	Proposed	38.09	2.670
		Naive	35.70	1.416
		Difference†	2.395	2.158
2	24.00	Proposed	32.21	2.132
		Naive	34.68	1.490
		Difference	−2.468	1.779
3	45.95	Proposed	27.44	2.653
		Naive	27.68	1.366
		Difference	−0.236	2.540
4	40.28	Proposed	24.42	1.687
		Naive	28.51	1.435
		Difference	−4.085	1.667

† difference = proposed − naive

patients with abnormal data, we focus on 1,271 patients with responses in four time intervals, (4,12], (12,20], (20,28], (28,36], denoted as y_1, y_2, y_3, y_4.

The longitudinal response $\boldsymbol{y} = (y_1, ..., y_4)^T$ has item nonresponse, as summarized in Table 5. The item nonresponse is due to adverse events, low-grade toxic reactions, the desire to seek other therapies, death, and some other reasons. Previous experiences from doctors and Cho et al. (2016) found that a steep decline in the CD4 cell count indicates the disease progression, and patients with low CD4 cell counts are more likely to miss the scheduled study visits as compared to patients with normal CD4. Therefore, nonresponse of the CD4 cell count is likely related to itself and is nonignorable (Cho et al. 2016; Yuan and Yin 2010).

We apply our proposed method in Sect. 2 to estimate $\mu_j = E(y_j)$, $j = 1, ..., 4$, with the always observed $z =$ the baseline CD4 measurement as the instrument described in (A1). Since z is the baseline CD4 cell count and $\boldsymbol{y}$ is the after-baseline CD4 cell count vector, based on the reason of nonresponse described previously, it is reasonable to assume that the item nonresponse of $\boldsymbol{y}$ is unrelated with the baseline z once we conditioned on $\boldsymbol{y}$, i.e., (A1) holds with z as an instrument. To apply the proposed method, we assume model (3) with $\boldsymbol{x} = z$, i.e., there is no other covariate.

The proposed estimates for $j = 1, ..., 4$ are shown in Table 6, together with their bootstrap standard errors with bootstrap size 200. For comparison, we also include in Table 6 the sample mean of observed values of y_j (naive estimate ignoring nonresponse), the differences between the proposed and naive estimates, and the bootstrap standard errors for differences.

From Table 6, the proposed estimates show a more serious decline in CD4 cell count over the time than naive estimates, although naive estimates also indicate the decline. Compared with 2 times the standard error, the difference between the proposed and naive estimates is significant at $j = 4$.

Acknowledgements We are grateful to the associate editor and two referees for comments and suggestions that led to improvements of the paper. Lyu Ni's research was supported by the Shanghai Sailing Program 22YF1411300. Jun Shao's research was supported by the National Natural Science Foundation of China Grant 11831008 and the U.S. National Science Foundation Grant DMS-1914411.

References

Cho, H., Hong, H. G., Kim, M. O. (2016). Efficient quantile marginal regression for longitudinal data with dropouts. *Biostatistics, 17,* 561–575.

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.

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

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica, 50,* 1029–1054.

Li, S., Shao, J. (2022). Nonignorable item nonresponse in panel data. *Statistical Theory and Related Fields, 6,* 58–71.

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

Rubin, D. B. (1976). Inference with missing data. *Biometrika, 63,* 581–592.

Shao, J., Tu, D. (1995). *The jackknife and bootstrap*. New York: Springer-Verlag.

Shao, J., Wang, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika, 103,* 175–187.

Shao, J., Zhang, J. (2015). A transformation approach in linear mixed-effect models with informative missing responses. *Biometrika, 102,* 107–119.

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

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.

Xu, L., Shao, J. (2009). Estimation in longitudinal or panel data models with random-effect-based missing responses. *Biometrics, 65,* 1175–1183.

Yuan, Y., Yin, G. (2010). Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics, 66,* 105–114.

Zhao, J., Shao, J. (2015). Semiparametric pseudo likelihoods in generalized linear models with nonignorable missing data. *Journal of American Statistical Association, 110,* 1577–1590.