Statistica Sinica **32** (2022), 271-292
doi:https://doi.org/10.5705/ss.202019.0243

FULL-SEMIPARAMETRIC-LIKELIHOOD-BASED INFERENCE FOR NON-IGNORABLE MISSING DATA

Yukun Liu, Pengfei Li and Jing Qin

East China Normal University, University of Waterloo and National Institute of Allergy and Infectious Diseases

Abstract: Most existing studies on missing-data problems focus on the ignorable missing case, where the missing probability depends only on observable quantities. By contrast, research on nonignorable missing data problems is quite limited. The main difficulty in solving such problems is that the missing probability and the regression likelihood function are tangled together in the likelihood presentation. Furthermore, the model parameters may not be identifiable, even under strong parametric model assumptions. In this paper, we discuss a semiparametric model for data with nonignorable missing responses, and propose a maximum full semiparametric likelihood estimation method. This method is an efficient combination of the parametric conditional likelihood and the marginal nonparametric biased sampling likelihood. We further show that the proposed estimators for the underlying parameters and the response mean are semiparametrically efficient. Extensive simulations and a real-data analysis demonstrate the advantage of the proposed method over competing methods.

Key words and phrases: Density ratio model, empirical likelihood, non-ignorable missing data.

1. Introduction

Missing data occur in many areas, including survey sampling, epidemiology, economics, sociology, and political science. Missing-data problems have been studied extensively during the last few decades. However, most research focuses on missing data that are ignorable or missing at random, in the sense that the missing probability or propensity score is a function only of the observed data (Little and Rubin (2002); Rubin (1987)).

Nonignorable missing or missing-not-at-random data occur if the propensity score depends on the missing data, even conditionally on the observed data. Let D be the missing indicator of the variable of interest Y associated with some covariate variables $\mathbf{X}$, and let $D = 1$ if Y is observed, and $D = 0$ otherwise.

Corresponding author: Yukun Liu, KLATASDS - MOE, School of Statistics, East China Normal University, Shanghai 200241, China. E-mail: ykliu@sfs.ecnu.edu.cn.

Nonignorable missing implies that the propensity score $\text{pr}(D = 1|\mathbf{x}, y) = \text{pr}(D = 1|\mathbf{X} = \mathbf{x}, Y = y)$ depends on y and possibly on $\mathbf{x}$. Inferences for nonignorable missing data are more challenging than those for ignorable missing data, for at least two reasons. First, the equality $\text{pr}(y|\mathbf{x}, D = 1) = \text{pr}(y|\mathbf{x}, D = 0)$, which holds for ignorable missing data, does not hold for nonignorable missing data. This implies that simply ignoring the missing data can lead to substantial selection bias (Groves, Presser and Dipko (2004)). Second, unlike the ignorable missing case, the propensity score and the regression likelihood function are tangled together in nonignorable missing-data problems, and hence cannot be estimated separately.

These challenges require new modeling strategies for nonignorable missing data. The most popular strategy is to make assumptions about $\text{pr}(D = 1|\mathbf{x}, y)$ and $\text{pr}(y|\mathbf{x})$, based on the selection model factorization $\text{pr}(y, D|\mathbf{x}) = \text{pr}(D|\mathbf{x}, y)\ \text{pr}(y|\mathbf{x})$ of Little and Rubin (2002). Parametric models (Greenless, Reece and Zieschang (1982); Baker and Laird (1988); Liu and Zhou (2010)) are at risk of model misspecification (Little (1985)), while completely nonparametric models suffer from the identifiability issue (Robins and Ritov (1997)). Attention has been paid to the case where one of these probabilities is parametric or semiparametric, and the other is left unspecified; see Tang, Little and Raghunathan (2003); Qin, Leung and Shao (2002); Chang and Kott (2008); Kott and Chang (2010), and Kim and Yu (2011). An alternative approach is to make parametric model assumptions on the observed Y, given $\mathbf{X}$ (Lee and Marsh (2000); Riddles, Kim and Im (2016)). An obvious advantage of this model over a completely parametric model for $\text{pr}(y|\mathbf{x})$ is that it can be checked using the available data.

Numerous estimation approaches for identifiable model parameters have been developed in recent years, including the pseudo-likelihood approaches (Tang, Little and Raghunathan (2003); Zhao and Shao (2015)), empirical likelihood method (Zhao, Zhao and Tang (2013); Tang, Zhao and Zhu (2014)), and generalized method of moments with an instrument variable (Wang, Shao and Kim (2014); Shao and Wang (2016); Shao (2018)); see Tang and Ju (2018) for a review of the most recent advances in dealing with nonignorable missing data. Under a parametric model for the observed Y given $\mathbf{X}$, Riddles, Kim and Im (2016), Morikawa, Kim and Kano (2017), and Morikawa and Kim (2016) proposed estimation equation methods based on the mean score equation of Louis (1982). However these approaches are either not efficient, or they suffer from the curse of dimensionality, as well as requiring a bandwidth selection. To avoid this dilemma, Ai, Linton and Zhang (2018) proposed a new estimation method based on the generalized method of moments with a diverging number of estimating equa-

tions. As the number of estimating equation increases, their estimator attains the semiparametric efficiency lower bound of Morikawa and Kim (2016). However, the constrained generalized method of moments may have numerical convergence problems, especially when some of the estimating equations are highly correlated.

In this study, we consider parametric models for both $\mathrm{pr}(y|\mathbf{x}, D = 1)$ and $\mathrm{pr}(D = 1|\mathbf{x}, y)$. In particular, we assume that $\mathrm{pr}(D = 1|\mathbf{x}, y)$ follows a logistic regression model,

$$\mathrm{pr}(D = 0|\mathbf{x}, y) = \frac{\exp(\alpha^* + \mathbf{x}^\top \beta + y\gamma)}{1 + \exp(\alpha^* + \mathbf{x}^\top \beta + y\gamma)}, \tag{1.1}$$

which is commonly used in practice. Under these assumptions, we find that the distribution pairs $\{\mathrm{pr}(y|\mathbf{x}, D = 1),\ \mathrm{pr}(y|\mathbf{x}, D = 0)\}$ and $\{\mathrm{pr}(\mathbf{x}|D = 1),\ \mathrm{pr}(\mathbf{x}|D = 0)\}$ satisfy two density ratio models (Anderson (1979, DRMs)); see Equations (2.4) and (2.5), which share some key unknown parameters. We give an easy-to-check condition to verify the identifiability of the model parameters. This condition is satisfied by many existing identification conditions, such as the existence of an instrument or ancillary variable (Wang, Shao and Kim (2014); Miao, Ding and Geng (2016)). For parameter estimation, the completely observed covariate data can be used to estimate the key unknown parameters, which can be further used to estimate $\mathrm{pr}(y|\mathbf{x}, D = 0)$, because $\mathrm{pr}(y|\mathbf{x}, D = 1)$ can be estimated directly using the conditional maximum likelihood method. These, together with the empirical distribution of D, lead to the estimation of the conditional density $\mathrm{pr}(y|\mathbf{x})$. As a result, we can consistently estimate the characteristics of Y.

Given the completely observed covariate data and the fact that $\{\mathrm{pr}(\mathbf{x}|D = 1),\ \mathrm{pr}(\mathbf{x}|D = 0)\}$ follows a DRM, we use the empirical likelihood (EL) of Owen (1988, 2001) to estimate the underlying parameters. The DRM-based EL has been demonstrated to be very flexible and efficient, and has attracted much attention in recent decades; see Qin and Zhang (1997), Chen and Liu (2013), Cai, Chen and Zidek (2017), and the references therein.

We show that the maximum EL estimators of the underlying parameters are asymptotically normal, and that the EL ratio for all the parameters follows an asymptotically central chi-square distribution. This makes it much more convenient to conduct hypothesis testing or to construct confidence intervals for these parameters. We propose a maximum likelihood estimator (MLE) for the marginal mean of the response variable, and establish its asymptotic normality. We further show that the proposed MLEs for all parameters attain the corresponding semiparametric efficiency lower bounds under parametric assumptions for the propensity score and the conditional density of Y, given $\mathbf{X}$ and $D = 1$.

Compared with existing methods, the proposed maximum semiparametric full likelihood approach has at least the following advantages:

1. It is able to identify the underlying parameters, regardless of whether an instrument variable exists, if the conditions in Proposition 1 are satisfied. The methods of Shao and Wang (2016), Riddles, Kim and Im (2016), Morikawa, Kim and Kano (2017), Morikawa and Kim (2016), and Ai, Linton and Zhang (2018) all require an instrument variable. Furthermore, it is able to produce consistent estimators for all of the model parameters, if they are identifiable. Extra information about the parameter γ in (1.1) is not needed.

2. It applies to data of any dimension, and is free of bandwidth selection. The methods of Kim and Yu (2011), Shao and Wang (2016), Morikawa and Kim (2016), and Morikawa, Kim and Kano (2017) all suffer from the curse of dimensionality and bandwidth selection, and may not work well for multivariate covariates. The method of Ai, Linton and Zhang (2018) has an increasing calculation burden as the number of estimating equation increases.

3. Existing methods handling nonignorable missing-data problems under semiparametric setups are mainly based on estimation equations, and may not be the most efficient, in general. Because full likelihood approaches are, in general, the most efficient, the proposed maximum semiparametric full likelihood approach is expected to outperform existing methods. Even though Morikawa and Kim (2016) calculated the semiparametric efficiency lower bound using only a specification of the propensity score, their lower bound is not achievable unless the conditional density of Y given $(\mathbf{X}, D = 1)$ is fully specified. Here we show that with the knowledge of $\mathrm{pr}(y|\mathbf{x}, D = 1)$, the method of Morikawa and Kim (2016) is no longer optimal. Our new lower bound is lower than theirs.

4. Our method is also applicable to data collected retrospectively. For example, when the number of nonresponse individuals (with $D = 0$) is large, we can randomly select some covariate $\mathbf{x}$ from them to save costs. Using these and the fully observed data, our method still provides a valid inference about the underlying population. In contrast, existing methods may produce biased estimators because they are designed for prospective data.

The rest of this paper is organized as follows. In Section 2, we introduce the proposed model, show its equivalence to two DRMs, and provide sufficient

conditions for the identifiability of the model parameters. Section 3 presents the proposed semiparametric DRM-based EL method and the resulting MLEs for the underlying parameters and the mean of the response variable. Their asymptotic normalities and semiparametric efficiencies are also established. Section 4 reports extensive simulation results. A real-life set of data is analyzed for illustration in Section 5. All technical details are given in the Supplementary Material.

2. Model and Its Identifiability

2.1. Model setup

Suppose $\{(y_i, \mathbf{x}_i, d_i), i = 1, \ldots, n\}$ are n independent and identically distributed (i.i.d.) copies of $(Y, \mathbf{X}, D)$, where the covariates $\mathbf{x}_i$ are always observed, and y_i is observed if and only if $d_i = 1$. We assume that the missing probability satisfies the logistic regression model in (1.1), that is,

$$\mathrm{pr}(D = 0|\mathbf{x}, y) = \frac{\exp(\alpha^* + \mathbf{x}^\top \beta + y\gamma)}{1 + \exp(\alpha^* + \mathbf{x}^\top \beta + y\gamma)}.$$

The parameter γ is called the tilting parameter (Kim and Yu (2011)). It quantifies the extent to which the model departs from ignorable missing, where $\gamma = 0$ corresponds to the ignorable missing-data case. We are interested in estimating the underlying parameters $(\alpha^*, \beta, \gamma)$ and the marginal mean μ of Y.

Based on the observed data, the full likelihood is

$$\prod_{i=1}^{n} \left[\{\mathrm{pr}(D = 1|\mathbf{x}_i, y_i)\mathrm{pr}(y_i, \mathbf{x}_i)\}^{d_i} \left\{ \int \mathrm{pr}(D = 0|\mathbf{x}_i, y)\mathrm{pr}(y, \mathbf{x}_i)dy \right\}^{1-d_i} \right]. \quad (2.1)$$

Unlike the case of ignorable missing, here, $\mathrm{pr}(D = 1|\mathbf{x}, y)$ and $\mathrm{pr}(y, \mathbf{x})$ cannot be separated, and hence cannot be separately estimated. To make an inference based on the full likelihood, one may postulate parametric assumptions on $\mathrm{pr}(D = 1|y, \mathbf{x})$ and $\mathrm{pr}(y|\mathbf{x})$, which are sensitive to model misspecification (Little (1985); Kenward and Molenberghs (1988)).

We solve this problem using an alternative method. The logistic regression model (1.1) is equivalent to the two-sample DRM (Qin and Zhang (1997))

$$\mathrm{pr}(\mathbf{x}, y|D = 0) = \exp(\alpha + \mathbf{x}^\top \beta + y\gamma)\mathrm{pr}(\mathbf{x}, y|D = 1), \quad (2.2)$$

where $\alpha = \alpha^* + \log\{\eta/(1 - \eta)\}$ and $\eta = \mathrm{pr}(D = 1)$ is the probability of being observed. Clearly, η can be consistently estimated from the data, and is therefore identifiable. Then, the identifiability of α^* is equivalent to that of α.

Integrating out y, we have

$$\mathrm{pr}(\mathbf{x}|D=0) = \exp(\alpha + \mathbf{x}^\top\beta)\mathrm{pr}(\mathbf{x}|D=1)\int \exp(y\gamma)\mathrm{pr}(y|\mathbf{x}, D=1)dy.$$

Therefore, the conditional densities of $Y = y$ given $(\mathbf{X} = \mathbf{x}, D = 0)$ and given $(\mathbf{X} = \mathbf{x}, D = 1)$ satisfy

$$\mathrm{pr}(y|\mathbf{x}, D=0) = \frac{\mathrm{pr}(\mathbf{x}, y|D=0)}{\mathrm{pr}(\mathbf{x}|D=0)} = \frac{\exp(y\gamma)\mathrm{pr}(y|\mathbf{x}, D=1)}{\int \exp(y\gamma)\mathrm{pr}(y|\mathbf{x}, D=1)dy}. \tag{2.3}$$

Although $\mathrm{pr}(y|\mathbf{x}, D = 1)$ is directly estimable from the observed $(y_i, \mathbf{x}_i)$ with $d_i = 1$, it is impossible to estimate $\mathrm{pr}(y|\mathbf{x}, D = 0)$, because γ is unknown, in general. As a result, the conditional approach is not viable, as demonstrated by Kim and Yu (2011), who rely on external data to identify γ. In practical applications, however, external data are often unavailable, which makes estimating γ impossible.

Fortunately, the marginal information on $(\mathbf{x}_i, d_i)$ can help to identify γ, which solves the identifiability problem in nonignorable missing-data problems. Because $(y_i, \mathbf{x}_i)$ with $d_i = 1$ are available, without loss of generality, we can postulate a parametric model $f(y|\mathbf{x}, \xi)$ for $\mathrm{pr}(y|\mathbf{x}, D = 1)$ with an identifiable parameter ξ. This parameter can be estimated consistently from the directly observed data. This parametric model and Equation (2.3) imply two DRMs:

$$\mathrm{pr}(y|\mathbf{x}, D=0) = \exp\{\gamma y - c(\mathbf{x}, \gamma, \xi)\}f(y|\mathbf{x}, \xi), \tag{2.4}$$

$$\mathrm{pr}(\mathbf{x}|D=0) = \exp\{\alpha + \mathbf{x}^\top\beta + c(\mathbf{x}, \gamma, \xi)\}\mathrm{pr}(\mathbf{x}|D=1), \tag{2.5}$$

where

$$c(\mathbf{x}, \gamma, \xi) = \ln\left\{\int \exp(y\gamma)f(y|\mathbf{x}, \xi)dy\right\}. \tag{2.6}$$

Equations (2.4)–(2.6) are the foundation of our inference method. Note that the second DRM involves all the underlying parameters in the model, and is dependent only on $\mathrm{pr}(\mathbf{x}|D = 0)$ and $\mathrm{pr}(\mathbf{x}|D = 1)$. Because the $(\mathbf{x}_i, d_i)$'s with $d_i = 0$ or 1 are not subject to missingness, the parameters can be estimated consistently using their maximum DRM-based EL estimators (Qin and Zhang (1997)), provided that they are identifiable.

2.2. Model identifiability

Miao, Ding and Geng (2016) pointed out that even under full parametric models for $\mathrm{pr}(D = 1|\mathbf{x}, y)$ and $\mathrm{pr}(y|\mathbf{x})$, the underlying model parameters may not

be identifiable. This phenomenon also arises under Model (2.5), where even when $\mathrm{pr}(\mathbf{x}|D=1)$ is completely known, the model parameters may not be identifiable. We present a simple-to-check sufficient condition for the identifiability of the underlying parameters in (2.5). We have assumed that ξ is identifiable. Hence, we focus here on the identifiability of the parameters α, β, and γ. Given the data $\{(\mathbf{x}_i, d_i), i = 1, \ldots, n\}$, the conditional density functions $\mathrm{pr}(\mathbf{x}|D=0)$ and $\mathrm{pr}(\mathbf{x}|D=1)$ are clearly identifiable and can be estimated consistently using, for example, the kernel method. The log ratio $\log\{\mathrm{pr}(\mathbf{x}|D=0)/\mathrm{pr}(\mathbf{x}|D=1)\}$ is also identifiable. Because

$$\log\left\{\frac{\mathrm{pr}(\mathbf{x}|D=0)}{\mathrm{pr}(\mathbf{x}|D=1)}\right\} = \alpha + \mathbf{x}^\top\beta + c(\mathbf{x}, \gamma, \xi),$$

the model identification is equivalent to identifying the parameters α, β, and γ in $\alpha + \mathbf{x}^\top\beta + c(\mathbf{x}, \gamma, \xi)$.

Proposition 1. *Let S be the common support of $\mathrm{pr}(\mathbf{x}|D=0)$ and $\mathrm{pr}(\mathbf{x}|D=1)$, and $\Omega = \{h(\mathbf{x}) : S \mapsto \mathbb{R} \mid \exists(\alpha, \beta, \gamma)$, such that $h(\mathbf{x}) = \alpha + \mathbf{x}^\top\beta + c(\mathbf{x}, \gamma, \xi)\ \forall\ \mathbf{x} \in S\}$. If for any $h(\mathbf{x}) \in \Omega$, there exists a unique (α, β, γ) such that $h(\mathbf{x}) = \alpha + \mathbf{x}^\top\beta + c(\mathbf{x}, \gamma, \xi)$, then (α, β, γ) is identifiable.*

Next, we apply the above proposition to some special cases. We need the concept of an instrument variable, which can be helpful to identify γ. Suppose $\mathbf{x}$ can be written as $\mathbf{x} = (z, u^\top)^\top$. If

$$\mathrm{pr}(D=0|z,u,y) = \mathrm{pr}(D=0|u,y) = \frac{\exp(\alpha^* + u^\top\beta + y\gamma)}{1+\exp(\alpha^* + u^\top\beta + y\gamma)}$$

and $\mathrm{pr}(y|\mathbf{x}) = \mathrm{pr}(y|z,u)$ depends on z and possibly on u, then z is an instrument variable. That is, an instrument variable is defined as a covariate that does not affect the missingness, but may affect the conditional distribution of the response variable.

With the above preparation and Proposition 1, we find that (α, β, γ) is identifiable in the following two cases.

Corollary 1. *Suppose the logistic regression model in (1.1) holds, and that the density function of Y given $(\mathbf{X} = \mathbf{x}, D = 1)$ is $f(y|\mathbf{x}, \xi)$. (a) If there exists an instrument variable z in $\mathbf{x}$, then (α, β, γ) is identifiable. (b) Assume that the set S in Proposition 1 contains an open set, and $c(\mathbf{x}, \gamma, \xi)$ can be expressed as $c(\mathbf{x}, \gamma, \xi) = \sum_{i=1}^{k} a_i(\gamma)g_i(\mathbf{x}) + a_{k+1}(\gamma) + \mathbf{x}^\top a_{k+2}(\gamma)$, for some positive integer k and continuous functions $a_i(\gamma)$ $(i = 1, \ldots, k+2)$ and $g_i(\mathbf{x})$ $(i = 1, \ldots, k)$, where $1, \mathbf{x}, g_1(\mathbf{x}), \ldots, g_k(\mathbf{x})$ are linearly independent, and $a_j(\gamma)$ $(j = 1, \ldots, k)$ are not*

equal to the zero functions. If $\big(a_1(\gamma_1),\ldots,a_k(\gamma_1)\big) \neq \big(a_1(\gamma_2),\ldots,a_k(\gamma_2)\big)$*, for any* $\gamma_1 \neq \gamma_2$*, then* (α,β,γ) *is identifiable.*

As an application of the above results, we consider the normal model in which $f(y|\mathbf{x},\xi)$ is the density function of $N\big(\mu(\mathbf{x},\xi),\sigma^2(\mathbf{x},\xi)\big)$. Direct calculations give $c(\mathbf{x},\gamma,\xi) = \gamma\mu(\mathbf{x},\xi) + 0.5\gamma^2\sigma^2(\mathbf{x},\xi)$. Furthermore, assume $\mu(\mathbf{x},\xi) = \mathbf{x}^\top b_1(\xi) + b_2(\xi)\mathbf{x}^\top\mathbf{x}$ and $\sigma^2(\mathbf{x},\xi) = \exp\{b_3(\xi) + \mathbf{x}^\top b_4(\xi)\}$ for nonzero functions $b_i(\xi)$. We have the following observations:

(I) If $b_2(\xi) \neq 0$, then according to Corollary 1, (α,β,γ) is identifiable.

(II) If $b_2(\xi) = 0$ and $b_4(\xi) = 0$, then

$$\alpha + \mathbf{x}^\top\beta + c(\mathbf{x},\gamma,\xi) = \alpha + 0.5\gamma^2\exp\{b_3(\xi)\} + \mathbf{x}^\top\{\beta + \gamma b_1(\xi)\},$$

which, together with Lemma 1, implies that (α,β,γ) is not identifiable.

(III) If $b_2(\xi) = 0$ and $b_4(\xi) \neq 0$, then

$$\alpha + \mathbf{x}^\top\beta + c(\mathbf{x},\gamma,\xi) = \alpha + \mathbf{x}^\top\{\beta + \gamma b_1(\xi)\} + 0.5\gamma^2\exp\{b_3(\xi) + \mathbf{x}^\top b_4(\xi)\}.$$

Furthermore, if $\gamma = 0$, then Proposition 1 implies that (α,β,γ) is identifiable. Otherwise, (α,β,γ) is not identifiable.

3. Semiparametric EL Inference

3.1. EL

Suppose there are n_1 completely observed data and n_2 partially observed data. Without loss of generality, we assume that $d_i = 1$, for $i = 1,\ldots,n_1$, and $d_i = 0$, for $i = n_1+1,\ldots,n$. The full likelihood in (2.1) can be written as

$$\prod_{i=1}^{n_1}\{\mathrm{pr}(y_i|\mathbf{x}_i,D=1)\mathrm{pr}(\mathbf{x}_i|D=1)\mathrm{pr}(D=1)\}\cdot\prod_{i=n_1+1}^{n}\{\mathrm{pr}(\mathbf{x}_i|D=0)\mathrm{pr}(D=0)\}.$$

Let $\theta = (\alpha,\beta^\top,\gamma,\xi^\top)^\top$ and $t(\mathbf{x},\theta) = \alpha + \mathbf{x}^\top\beta + c(\mathbf{x},\gamma,\xi)$. Because $\mathrm{pr}(y|\mathbf{x},D=1) = f(y|\mathbf{x},\xi)$ by assumption, it follows from $\eta = \mathrm{pr}(D=1)$ and Equation (2.5) that the full log-likelihood is $\tilde{\ell} = \ell_1(\eta) + \tilde{\ell}_2$, where

$$\ell_1(\eta) = n_1\log(\eta) + (n-n_1)\log(1-\eta)$$

is the marginal likelihood based on d_i, and

$$\tilde{\ell}_2 = \sum_{i=1}^{n_1} \log\{f(y_i|\mathbf{x}_i,\xi)\} + \sum_{i=n_1+1}^{n} t(\mathbf{x}_i,\theta) + \sum_{i=1}^{n} \log\{\mathrm{pr}(\mathbf{x}_i|D=1)\}$$

is the conditional likelihood given d_i.

We leave the conditional density $\mathrm{pr}(\mathbf{x}|D=1)$ completely unspecified, and handle it using the EL method of Owen (1988, 1990). Let $p_i = \mathrm{pr}(\mathbf{x}_i|D=1) = dF(\mathbf{x}_i|D=1)$, where $F(\mathbf{x}|D=1)$ is the cumulative distribution function corresponding to the density $\mathrm{pr}(\mathbf{x}|D=1)$. Following the principle of EL, $\tilde{\ell}_2$ becomes an empirical log-likelihood

$$\tilde{\ell}_2 = \sum_{i=1}^{n_1} \log\{f(y_i|\mathbf{x}_i,\xi)\} + \sum_{i=n_1+1}^{n} t(\mathbf{x}_i,\theta) + \sum_{i=1}^{n} \log(p_i),$$

where p_i is subject to the constraints

$$p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i[\exp\{t(\mathbf{x}_i,\theta)\} - 1] = 0.$$

Maximizing $\tilde{\ell}_2$ with respect to p_i, we have

$$p_i = \frac{1}{n}\frac{1}{1+\lambda[\exp\{t(\mathbf{x}_i,\theta)\}-1]}, \tag{3.1}$$

where λ is the solution to

$$\sum_{i=1}^{n} \frac{\exp\{t(\mathbf{x}_i,\theta)\}-1}{1+\lambda[\exp\{t(\mathbf{x}_i,\theta)\}-1]} = 0. \tag{3.2}$$

Substituting p_i into $\tilde{\ell}_2$ leads to the profile log-likelihood of θ,

$$\ell_2(\theta) = \sum_{i=1}^{n_1} \log\{f(y_i|\mathbf{x}_i,\xi)\} + \sum_{i=n_1+1}^{n} t(\mathbf{x}_i,\theta) - \sum_{i=1}^{n} \log\big\{1+\lambda[\exp\{t(\mathbf{x}_i,\theta)\}-1]\big\}.$$

The profile log-likelihood of (η,θ) is then defined as

$$\ell(\eta,\theta) = \ell_1(\eta) + \ell_2(\theta). \tag{3.3}$$

3.2. Estimation of the underlying parameters

From the profile log-likelihood of (η,θ) in (3.3), the MLE of (η,θ) is

$$(\hat{\eta},\hat{\theta}) = \underset{\eta,\theta}{\mathrm{argmax}}\, \ell(\eta,\theta).$$

Equivalently, $\hat{\eta}$ maximizes $\ell_1(\eta)$, which gives $\hat{\eta} = n_1/n$, and $\hat{\theta} = (\hat{\alpha}, \hat{\beta}^\top, \hat{\gamma}, \hat{\xi}^\top)^\top = \arg\max_\theta \ell_2(\theta)$. The likelihood ratio function of θ is defined as

$$R(\theta) = 2\{\max_{\eta,\theta} \ell(\eta, \theta) - \max_{\eta} \ell(\eta, \theta)\} = 2\{\ell_2(\hat{\theta}) - \ell_2(\theta)\}.$$

Next, we study the large-sample properties of the MLE and the likelihood ratio. Denote the truth of (η, θ) by (θ_0, η_0), with $\theta_0 = (\alpha_0, \beta_0^\top, \gamma_0, \xi_0^\top)^\top$ and $\eta_0 \in (0, 1)$. Define

$$\pi(\mathbf{x}; \theta, \eta) = \frac{(1-\eta)\exp\{t(\mathbf{x}, \theta)\}}{\eta + (1-\eta)\exp\{t(\mathbf{x}, \theta)\}},$$

which we abbreviate as $\pi(\mathbf{x}) = \pi(\mathbf{x}; \theta_0, \eta_0)$. Let d_θ denote the dimension of θ, and let $\mathbf{e}_1$ be a $d_\theta \times 1$ vector, with the first component being one, and the remaining components zero. Finally, define

$$V = \mathbb{E}[\{1 - \pi(\mathbf{X})\}\pi(\mathbf{X})\{\nabla_\theta t(\mathbf{X}, \theta)\}^{\otimes 2}] + \mathbb{E}[DI_e\{\nabla_\xi f(Y|\mathbf{X}, \xi)\}^{\otimes 2} I_e^\top], \quad (3.4)$$

where ∇_θ is the differentiation operator with respect to θ, $I_e^\top = (0_{d_\xi \times (2+d_\beta)}, I_{d_\xi \times d_\xi})$, and $B^{\otimes 2} = BB^\top$ for any matrix or vector B.

Theorem 1. *Assume Conditions* A1–A4 *in the Supplementary Material hold. Suppose that the logistic regression model in* (1.1) *holds, with* $(\alpha_0, \beta_0, \gamma_0)$ *in place of* (α, β, γ)*, and that the density function of* Y *given* $(\mathbf{X} = \mathbf{x}, D = 1)$ *is* $f(y|\mathbf{x}, \xi_0)$*. Furthermore, assume that* θ *is identifiable. Then, as* $n \to \infty$*, (1)* $\sqrt{n}(\hat{\theta} - \theta_0) \to N\big(0,\ V^{-1} - \{\eta_0(1-\eta_0)\}^{-1}\mathbf{e}_1\mathbf{e}_1^\top\big)$ *in distribution, with* V *defined in* (3.4)*; (2)* $R(\theta_0) \to \chi^2_{d_\theta}$ *in distribution.*

Theorem 1 implies that the MLEs of all the parameters are asymptotically normal. The likelihood ratio for the parameters follows a central chi-square limiting distribution, which makes the resulting hypothesis testing or interval estimation about θ very convenient. Although the proposed approach is developed based on prospective data, we emphasize that it can also apply to data collected retrospectively. This is because the subsequent inferences are mainly based on ℓ_2 or, equivalently,

$$\tilde{\ell}_2 = \log\left[\prod_{i=1}^{n_1}\{\mathrm{pr}(y_i, \mathbf{x}_i|D=1)\}\prod_{i=n_1+1}^{n}\{\mathrm{pr}(\mathbf{x}_i|D=0)\right],$$

which is actually a retrospective log-likelihood. If $\eta = \mathrm{pr}(D=1)$ or $\hat{\eta}$ is available, based on data collected retrospectively, the proposed approach can still make a valid inference.

Given the MLE of all the underlying parameters, we are able to construct the MLE of the population mean μ of the response Y. Under our model, μ depends not only on the underlying parameters θ, but also on $\mathrm{pr}(\mathbf{x}|D=1)$, or the corresponding cumulative distribution function $F(\mathbf{x}|D=1)$. Using the MLEs $\hat{\theta}$ and $\hat{\eta} = n_1/n$, we show in the Supplementary Material that $\hat{\lambda} = n_2/n$, where $\hat{\lambda}$ satisfies (3.2), with $\hat{\theta}$ in place of θ. From (3.1), the MLE of p_i is

$$\hat{p}_i = \frac{1}{n}\frac{1}{1+(n_2/n)[\exp\{t(\mathbf{x}_i,\hat{\theta})\}-1]} = \frac{1}{n}\frac{1}{\hat{\eta}+(1-\hat{\eta})\exp\{t(\mathbf{x}_i,\hat{\theta})\}}.$$

Accordingly, the MLE of $F(\mathbf{x}|D=1)$ is $\hat{F}(\mathbf{x}|D=1) = \sum_{i=1}^{n_1}\hat{p}_i I(\mathbf{x}_i \le \mathbf{x})$, where for two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, $\mathbf{x}_1 \le \mathbf{x}_2$ implies that the inequality holds, elementwise.

3.3. Estimation of the response mean

To obtain the MLE of the response mean μ, we write μ in terms of the underlying parameters η, θ, and $F(\mathbf{x}|D=1)$, as follows:

$$\begin{aligned}
\mu &= \int_y\int_{\mathbf{X}} y\mathrm{pr}(y|\mathbf{x},D=1)\mathrm{pr}(\mathbf{x}|D=1)\mathrm{pr}(D=1)d\mathbf{x}dy \\
&\quad + \int_y\int_{\mathbf{X}} y\mathrm{pr}(y|\mathbf{x},D=0)\mathrm{pr}(\mathbf{x}|D=0)\mathrm{pr}(D=0)d\mathbf{x}dy \\
&= \int_y\int_{\mathbf{X}} y\mathrm{pr}(y|\mathbf{x},D=1)\mathrm{pr}(\mathbf{x}|D=1)\eta d\mathbf{x}dy \\
&\quad + \int_y\int_{\mathbf{X}} y\exp(\alpha+\mathbf{x}^\top\beta+\gamma y)\mathrm{pr}(y|\mathbf{x},D=1)\mathrm{pr}(\mathbf{x}|D=1)(1-\eta)d\mathbf{x}dy \\
&= \int_{\mathbf{X}}\left[\int_y y\{\eta+(1-\eta)\exp(\alpha+\mathbf{x}^\top\beta+\gamma y)\}f(y|\mathbf{x},\xi)dy\right]dF(\mathbf{x}|D=1),
\end{aligned}$$

where in the last step, we replace $\mathrm{pr}(y|\mathbf{x},D=1)$ and $\mathrm{pr}(\mathbf{x}|D=1)d\mathbf{x}$ with $f(y|\mathbf{x},\xi)$ and $dF(\mathbf{x}|D=1)$, respectively. Then, the MLE of μ is

$$\begin{aligned}
\hat{\mu} &= \sum_{i=1}^n \hat{p}_i\left[\int_y y\{\hat{\eta}+(1-\hat{\eta})\exp(\hat{\alpha}+\mathbf{x}_i^\top\hat{\beta}+\hat{\gamma}y)\}f(y|\mathbf{x}_i,\hat{\xi})dy\right] \\
&= \frac{1}{n}\sum_{i=1}^n \frac{\int_y y\{\hat{\eta}+(1-\hat{\eta})\exp(\hat{\alpha}+\mathbf{x}_i^\top\hat{\beta}+\hat{\gamma}y)\}f(y|\mathbf{x}_i,\hat{\xi})dy}{\hat{\eta}+(1-\hat{\eta})\exp\{t(\mathbf{x}_i,\hat{\theta})\}}. \qquad (3.5)
\end{aligned}$$

We use the normal model as an illustrating example: $f(y|\mathbf{x},\xi)$ is chosen as the density function of $N\big(\mu(\mathbf{x},\xi),\sigma^2(\mathbf{x},\xi)\big)$. In this example, the proposed mean

estimator in (3.5) becomes

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\eta}\hat{\mu}_i + (1-\hat{\eta})(\hat{\mu}_i + \hat{\gamma}\hat{\sigma}_i^2)\exp(\hat{\alpha} + \mathbf{x}_i^\top \hat{\beta} + \hat{\mu}_i\hat{\gamma} + 0.5\hat{\gamma}^2\hat{\sigma}_i^2)}{\hat{\eta} + (1-\hat{\eta})\exp(\hat{\alpha} + \mathbf{x}_i^\top \hat{\beta} + \hat{\mu}_i\hat{\gamma} + 0.5\hat{\gamma}^2\hat{\sigma}_i^2)}, \quad (3.6)$$

where $\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\xi})$ and $\hat{\sigma}_i^2 = \sigma^2(\mathbf{x}_i, \hat{\xi})$.

The next theorem establishes the asymptotic normality of the proposed estimator $\hat{\mu}$ in (3.5).

Theorem 2. *Under the conditions of Theorem* 1*, as n goes to infinity,* $\sqrt{n}(\hat{\mu} - \mu) \to N(0, \sigma^2)$ *in distribution, where* $\sigma^2 = \mathbb{V}\mathrm{ar}\{K(\mathbf{X}; \theta_0, \eta_0)\} + A^\top V^{-1} A$*, with*

$$K(\mathbf{x}; \theta, \eta) = \frac{\int y\{\eta + (1-\eta)\exp(\alpha + \mathbf{x}^\top\beta + \gamma y)\} f(y|\mathbf{x}, \xi) dy}{\eta + (1-\eta)\exp\{\alpha + \mathbf{x}^\top\beta + c(\mathbf{x}, \gamma, \xi)\}}$$

and $A = \mathbb{E}\{\nabla_\theta K(\mathbf{X}; \theta_0, \eta_0)\}$.

When Wald-type intervals are constructed for μ based on Theorem 2, we need a consistent estimator of σ^2, which can be constructed based on consistent estimators of A, $\mathbb{V}\mathrm{ar}\{K(\mathbf{X}; \theta_0, \eta_0)\}$, and V. Reasonable estimators for these three quantities are $\hat{A} = n^{-1}\sum_{i=1}^n \nabla_\theta K(\mathbf{x}_i; \hat{\theta}, \hat{\eta})$,

$$\widehat{\mathbb{V}\mathrm{ar}}\{K(\mathbf{X}; \theta_0, \eta_0)\} = n^{-1}\sum_{i=1}^{n}\{K(\mathbf{X}; \hat{\theta}, \hat{\eta})\}^2 - \left\{n^{-1}\sum_{i=1}^{n} K(\mathbf{X}; \hat{\theta}, \hat{\eta})\right\}^2,$$

and

$$\hat{V} = n^{-1}\sum_{i=1}^{n}[\{1 - \pi(\mathbf{x}_i, \hat{\theta}, \hat{\eta})\}\pi(\mathbf{x}_i, \hat{\theta}, \hat{\eta})\{\nabla_\theta t(\mathbf{x}_i, \hat{\theta})\}^{\otimes 2} + d_i I_e\{\nabla_\xi f(y_i|\mathbf{x}_i, \hat{\xi})\}^{\otimes 2} I_e^\top].$$

These estimators are consistent because $(\hat{\theta}, \hat{\eta})$ is consistent and K is smooth in all its arguments. Consequently, a consistent estimator of σ^2 is

$$\hat{\sigma}^2 = \widehat{\mathbb{V}\mathrm{ar}}\{K(\mathbf{X}; \theta_0, \eta_0)\} + \hat{A}^\top \hat{V}^{-1}\hat{A}. \quad (3.7)$$

3.4. Semiparametric efficiency

We make the same model assumptions as Riddles, Kim and Im (2016): the logistic model in (1.1) for the propensity score, and a parametric model $f(y|\mathbf{x}, \xi)$ for $\mathrm{pr}(y|\mathbf{x}, D = 1)$, leaving $\mathrm{pr}(\mathbf{x}|D = 1)$ completely unspecified. Therefore our model setup is semiparametric. Next, we show that the estimators $(\hat{\theta}, \hat{\eta})$ and $\hat{\mu}$, which are built on the above semiparametric model, are semiparametrically efficient.

Theorem 3. *Under the conditions of Theorem* 1*, the MLEs* $(\hat{\theta}, \hat{\eta})$ *and* $\hat{\mu}$ *are both semiparametrically efficient, in sense that their asymptotic variances attain the corresponding semiparametric efficiency lower bounds.*

We make some comments on Theorem 3 and the results in Riddles, Kim and Im (2016), Morikawa and Kim (2016), and Ai, Linton and Zhang (2018). Note that the Riddles, Kim and Im (2016) estimator was constructed under the same model assumptions as ours. Theorem 3 implies that the asymptotic variance of their mean estimator is no less than σ^2, which is the asymptotic variance of the MLE $\hat{\mu}$, as well as the semiparametric efficiency lower bound for estimating μ.

When only the parametric propensity score assumption is made, Morikawa and Kim (2016) derived the semiparametric efficiency lower bound for the parameter of interest, such as the response mean, and proposed two adaptive estimators with asymptotic variances that attain the lower bound. Ai, Linton and Zhang (2018) proposed an estimation method based on the generalized method of moments, showing that as the number of moments increases appropriately, their estimator also attains the lower bound of Morikawa and Kim (2016). According to Tsiatis (2006), the semiparametric efficiency lower bound is equal to the supremum of the asymptotic variances of the MLEs under all parametric submodels. Because the model assumptions in Morikawa and Kim (2016) are weaker than ours, the set of all parametric submodels considered in Morikawa and Kim (2016) contains all parametric submodels considered here. Consequently, when the parameter of interest is the response mean, the semiparametric efficiency lower bound of Morikawa and Kim (2016) is no less than σ^2. Hence, the asymptotic variances of the two Morikawa and Kim (2016) adaptive estimators and the Ai, Linton and Zhang (2018) estimator are no less than that of our estimator $\hat{\mu}$.

3.5. Model checking

Based on the completely observed data $\{(y_i, \mathbf{x}_i, d_i = 1), i = 1, \ldots, n_1\}$, we can directly examine the correctness of the model assumption $\text{pr}(y|\mathbf{x}, D = 1) = f(y|\mathbf{x}, \xi)$ using a residual analysis. For example, the goal of checking the normal model assumed in the real-data analysis in Section 5 can be achieved using popular normality tests, such as the Shapiro–Wilk test based on the residuals. We can perform model diagnostics using the Cox and Snell (1968) general residuals for other types of continuous responses, and the Yang (2019) surrogate empirical residual distribution function for discrete responses.

Another question about the proposed model is the reliability of the parametric model assumption on the propensity score $\text{pr}(D = 1|\mathbf{x}, y)$. Because we do

not observe y_i for $\{(\mathbf{x}_i, d_i = 0), i = n_1 + 1, \ldots, n\}$, we do not have direct data to check this. However, the question can be answered indirectly by testing the goodness-of-fit of the DRM (2.5). The latter problem has been studied by many researchers and can be solved using the tests of Qin and Zhang (1997), Cheng and Chu (2004), Bondell (2007), and others.

4. Simulation

4.1. Setup

We carry out simulations to investigate the finite-sample performance of the proposed estimator for the population mean of the response. We compare the proposed mean estimator $\hat{\mu}$ with four others: (1) the Morikawa and Kim (2016) adaptive estimator, with a correctly specified parametric form for $\mathrm{pr}(y|\mathbf{x}, D = 1)$, $\tilde{\mu}_t$; (2) the Morikawa and Kim (2016) adaptive estimator, without specifying a parametric form for $\mathrm{pr}(y|\mathbf{x}, D = 1)$, $\tilde{\mu}_{np}$; (3) the sample mean of the observed response, $\bar{y}_r$; and (4) the sample mean of all the responses when there are no missing data, $\bar{y}$. When $\mathrm{pr}(y|\mathbf{x}, D = 1)$ is correctly specified, Morikawa and Kim (2016) show that $\tilde{\mu}_t$ is more efficient than the estimator of Riddles, Kim and Im (2016), and that the Ai, Linton and Zhang (2018) estimator has the same asymptotic variance as $\tilde{\mu}_t$. Hence, the methods of Riddles, Kim and Im (2016) and Ai, Linton and Zhang (2018) are not included in the comparison. We also compare the proposed estimator of the unknown parameters in the missing probability model (1.1) with the adaptive estimators of Morikawa and Kim (2016). The results are summarized in Section S7 of the Supplementary Material.

We generate data from the following four examples.

Example 1. Let $\mathbf{x} = (z, u)^\top$, where u is a Bernoulli random variable with success probability 0.5, z follows the uniform distribution on $(-1, 1)$, and u and z are independent. We choose $\mathrm{pr}(D = 1|\mathbf{x}, y) = 1/\{1 + \exp(-1.7 - 0.4u + 0.5y)\}$, and set $\mathrm{pr}(y|\mathbf{x}, D = 1) = f(y|\mathbf{x}, \xi)$ to the density function of $N\big(\mu(\mathbf{x}), \sigma^2\big)$, where $\mu(\mathbf{x}) = \exp(0.5 - u + 1.5z)$ and $\sigma^2 = 1$ or 4.

Example 2. Let $\mathbf{x} = (z, u)^\top$, where $u \sim N(1, 1)$, $z \sim N(0, 1)$ and u and z are independent. We choose $\mathrm{pr}(D = 1|\mathbf{x}, y) = 1/\{1 + \exp(-1.7 - 0.4u + 0.5y)\}$, and set $\mathrm{pr}(y|\mathbf{x}, D = 1) = f(y|\mathbf{x}, \xi)$ to the density function of $N\big(\mu(\mathbf{x}), \sigma^2\big)$, where $\mu(\mathbf{x}) = 2.5 - u + 1.5z$ and $\sigma^2 = 1$ or 4.

Example 3. The covariate x follows $N(0, 1)$. We choose $\mathrm{pr}(D = 1|x, y) = 1/\{1 + \exp(-2.7 - 0.4x + 0.5y)\}$, and set $\mathrm{pr}(y|x, D = 1) = f(y|x, \xi)$ to the density function of $N\big(\mu(x), \sigma^2 e^{0.5x}\big)$, where $\mu(x) = 2 - x + x^2$ and $\sigma^2 = 1$ or $e^{0.7}$.

Table 1. True values of μ and the missing probability $1-\eta$ in Examples 1–4.

Example	σ^2	μ	$1-\eta$	Example	σ^2	μ	$1-\eta$
1	1	1.748	0.294	1	4	2.326	0.362
2	1	1.638	0.275	2	4	2.177	0.339
3	1	3.127	0.277	3	$e^{0.7}$	3.289	0.299
4	3	1.657	0.277	4	6	1.704	0.282

Example 4. The setup is the same as Example 2, except that $\mathrm{pr}(y|\mathbf{x}, D=1)$ is set to the density function of a normal mixture $0.95N\big(\mu(\mathbf{x}),1\big)+0.05N\big(\mu(\mathbf{x}),\sigma^2\big)$, where $\mu(\mathbf{x}) = 2.5 - u + 1.5z$ and $\sigma^2 = 3$ or 6.

Example 1 is Scenario 2 of Morikawa and Kim (2016), except that we consider $\sigma^2 = 1$ and 4, whereas Morikawa and Kim (2016) considered only $\sigma^2 = 1$. Example 2 represents the case where the mean function is a linear function of $\mathbf{x}$. Both Examples 1 and 2 have an instrument variable so the model parameters are identifiable. Example 3 represents the case of no instrument variable, but the model parameters are still identifiable. In Examples 1–3, the model for $\mathrm{pr}(y|\mathbf{x}, D=1)$ is correctly specified when implementing the proposed method. In Example 4, we choose $f(y|\mathbf{x},\xi)$ as the density function of $N\big(\mu(\mathbf{x}),\sigma^2\big)$ when implementing the proposed method, although the true density function for $\mathrm{pr}(y|\mathbf{x}, D=1)$ is a normal mixture. In this situation, $f(y|\mathbf{x},\xi)$ is a misspecified model for $\mathrm{pr}(y|\mathbf{x}, D=1)$. The true values of μ and the missing probability $1-\eta$ for the four examples are tabulated in Table 1.

4.2. Point estimation

In this section, we evaluate the performance of the five mean estimators in terms of the relative bias (RB) and the mean squared error (MSE). We set $n = 500$ and 2,000 for all four examples, and use 2,000 for the number of repetitions in all simulations. The simulation results are summarized in Table 2.

Note that we encountered some numerical problems when implementing the adaptive estimator of Morikawa and Kim (2016) $\tilde{\mu}_t$ in Example 1 with $n = 500, 2000$ and $\sigma^2 = 4$, in Example 2 with $n = 500$ and $\sigma^2 = 4$, in Example 3 with $n = 500$ and $\sigma^2 = e^{0.7}$, and in Example 4 with $n = 500, 2000$ and $\sigma^2 = 6$. Their algorithm either does not converge, or it produces too big (greater than five) or too small (less than zero) mean estimates. Throughout the simulation study, the performance of $\tilde{\mu}_t$ is evaluated based only on estimates between zero and five.

We first examine the results for Example 1. When $\sigma^2 = 1$, the relative biases of the proposed estimator and the two Morikawa and Kim (2016) adaptive

Table 2. Relative bias (RB; ×100) and mean squared error (MSE; ×100) of five estimates of μ.

n		$\hat{\mu}$	$\tilde{\mu}_t$	$\tilde{\mu}_{np}$	$\bar{y}_r$	$\bar{y}$	$\hat{\mu}$	$\tilde{\mu}_t$	$\tilde{\mu}_{np}$	$\bar{y}_r$	$\bar{y}$
		Example 1: $\sigma^2 = 1$					Example 1: $\sigma^2 = 4$				
500	RB	-0.12	-0.39	-1.31	-32.61	-0.19	0.35	-1.18	-8.45	-51.71	-0.19
500	MSE	0.93	0.98	1.04	33.10	0.81	4.00	7.17	7.18	146.23	1.78
2,000	RB	0.10	0.04	-0.33	-32.54	0.03	0.18	-0.11	-4.21	-51.57	0.01
2,000	MSE	0.22	0.24	0.24	32.49	0.19	0.98	1.24	1.91	144.24	0.44
		Example 2: $\sigma^2 = 1$					Example 2: $\sigma^2 = 4$				
500	RB	-0.15	-0.28	-4.49	-35.83	0.01	0.18	-0.68	-18.62	-56.27	0.05
500	MSE	1.09	1.12	1.62	35.49	0.93	3.97	5.68	19.45	152.19	1.90
2,000	RB	0.14	0.13	-2.85	-35.69	0.11	0.15	0.09	-15.01	-56.05	0.09
2,000	MSE	0.26	0.27	0.48	34.43	0.23	0.97	1.40	11.43	149.48	0.47
		Example 3: $\sigma^2 = 1$					Example 3: $\sigma^2 = e^{0.7}$				
500	RB	0.01	-0.24	-2.29	-18.70	0.06	0.02	-0.71	-3.08	-23.00	0.06
500	MSE	1.01	1.82	1.48	34.79	0.90	1.59	3.99	2.52	58.19	1.21
2,000	RB	0.05	-0.02	-1.20	-18.70	-0.04	0.05	-0.11	-1.67	-23.06	0.00
2,000	MSE	0.25	0.38	0.39	34.35	0.23	0.41	0.80	0.69	57.76	0.29
		Example 4: $\sigma^2 = 3$					Example 4: $\sigma^2 = 6$				
500	RB	-0.39	-0.29	-5.08	-36.78	-0.07	-1.48	-0.39	-6.97	-38.68	0.22
500	MSE	1.10	1.18	1.78	38.19	0.94	1.25	1.77	2.56	44.52	1.02
2,000	RB	-0.14	0.01	-3.52	-36.67	0.09	-1.50	0.02	-5.56	-38.84	0.14
2,000	MSE	0.27	0.30	0.61	37.16	0.23	0.36	0.53	1.19	44.05	0.26

estimators are all small. The proposed estimator has slightly smaller MSEs than those of the two adaptive estimators, the MSEs of which are quite close to each other. When σ^2 is increased to four, the relative biases of $\tilde{\mu}_{np}$ become much bigger. The proposed estimator has much smaller MSEs than those of the two adaptive estimators. The comparison between $\hat{\mu}$ and $\tilde{\mu}_t$ in Example 2 is similar to that for Example 1. For Example 2, compared with $\tilde{\mu}_t$, $\tilde{\mu}_{np}$ has much bigger relative biases and MSEs, especially for larger σ^2. Next, we examine the results for Example 3, in which there is no instrumental variable. The proposed estimator has small relative biases in all situations. Its MSEs are significantly smaller than those of the two adaptive estimators. For Example 4, although the model for $\text{pr}(y|\mathbf{x}, D = 1)$ is misspecified, the relative biases of $\hat{\mu}$ are still small, which shows that the proposed method is quite robust to model misspecification. The comparison between $\hat{\mu}$ and the two Morikawa and Kim (2016) adaptive estimators in Example 4 is similar to that in Example 2. Finally, as expected, $\bar{y}_r$ has large relative biases and the largest MSEs in all examples, whereas the ideal estimator $\bar{y}$ has small relative biases and the smallest MSEs in all situations. When σ^2

is small, the proposed estimator has almost the same performance as the ideal estimator $\bar{y}$, indicating that it is nearly optimal.

4.3. Interval estimation

This section compares the coverage of Wald confidence intervals based on $\hat{\mu}$, $\tilde{\mu}_t$, and $\bar{y}_r$. The nonparametric bootstrap method with 200 bootstrap samples is used to estimate the asymptotic variance for each of the three mean estimators. Although the variance estimator in (3.7) can be used in the Wald-type confidence intervals based on $\hat{\mu}$, its complicated form makes it more difficult to calculate than the bootstrap variance estimate. The bootstrap method is quite computationally intensive for $\tilde{\mu}_{np}$. For example, in Example 1, it takes around nine minutes and two hours to calculate the bootstrap variances for $\tilde{\mu}_{np}$ for a single replication when $n = 500$ and $n = 2{,}000$, respectively. Hence, we do not include it in the comparison. Again, the number of repetitions is 2,000 in all cases. The simulation results are summarized in Table 3.

In most cases, both Wald confidence intervals based on $\hat{\mu}$ and $\tilde{\mu}_t$ have very close and accurate coverage probabilities. The exceptions are Example 1 and Example 3 with the smaller sample size $n = 500$, and Example 4. For Example 1, both intervals have slight under-coverage, whereas for Example 3, the Wald confidence interval based on $\tilde{\mu}_t$ has much over-coverage, particularly when σ^2 is large. When the sample size n is increased to 2,000, both intervals have perfect coverage accuracy. For Example 4, the Wald confidence interval based on $\tilde{\mu}_t$ has under-coverage, especially when $n = 2{,}000$. The Wald confidence interval based on $\hat{\mu}$ has a similar problem when $\sigma^2 = 6$ and $n = 2{,}000$. This is probably the cost of the model misspecification. Note that for the Wald confidence interval based on $\tilde{\mu}_t$, the results for $\tilde{\mu}_t$ outside $[0, 5]$ or not convergent were not considered. In all cases, the Wald confidence interval based on $\bar{y}_r$ has unacceptable coverage accuracy, most probably because of the severe bias of $\bar{y}_r$. Overall, the Wald confidence interval based on the proposed estimator $\hat{\mu}$ is the most accurate and desirable among the three interval estimators under comparison.

5. An Application

We apply the proposed method to analyze human immunodeficiency virus (HIV) data from the AIDS Clinical Trials Group Protocol 175 (ACTG175) (Hammer et al. (1996); Zhang and Wang (2020)), in which $n = 2{,}139$ HIV-infected patients were enrolled. The patients were randomly divided into four arms, according to the regimen of treatment they received: (I) zidovudine monotherapy,

Table 3. Simulated coverage probabilities (%) of bootstrap Wald-type confidence intervals based on $\hat{\mu}$, $\tilde{\mu}_t$, and $\bar{y}_r$ in Examples 1–4.

Example	σ^2	n	$\hat{\mu}$	$\tilde{\mu}_t$	$\bar{y}_r$	Example	σ^2	n	$\hat{\mu}$	$\tilde{\mu}_t$	$\bar{y}_r$
1	1	500	93.6	94.4	0	1	4	500	95.1	95.7	1.0
1	1	2,000	95.3	95.1	0	1	4	2,000	94.7	94.2	0
2	1	500	94.5	94.7	0.1	2	4	500	95.2	95.3	0
2	1	2,000	95.1	95.2	0	2	4	2,000	95.4	95.2	0
3	1	500	94.9	96.0	0	3	$e^{0.7}$	500	95.7	97.5	0
3	1	2,000	95.0	94.7	0	3	$e^{0.7}$	2,000	94.8	95.5	0
4	3	500	95.7	94.4	0	4	6	500	94.8	93.6	0
4	3	2,000	95.1	93.9	0	4	6	2,000	92.7	92.1	0

(II) zidovudine + didanosine, (III) zidovudine + zalcitabine, and (IV) didanosine monotherapy. The data record many measurements from each patient, including his/her age (in years), weight (in kilograms), CD4 cell count at baseline (cd40), CD4 cell count at 20±5 weeks (cd420), CD4 cell count at 96±5 weeks (cd496), CD8 cell count at baseline (cd80), CD8 cell count at 20±5 weeks (cd820), and arm number (arms). The data are available from the `R` package `speff2trial`. The effectiveness of an HIV treatment can be assessed by monitoring the CD4 cell counts of HIV-positive patients: an increase indicates an improvement in the patients' health. An interesting problem is to determine the mean of the CD4 cell counts in each arm after the patients were treated for about 96 weeks. We take cd496 as the response variable Y, and take age, weight, cd40, cd420, cd80, and cd820 as covariates $X_1, \ldots, X_6$, respectively. Owing to the end of the trial or loss to follow-up, 62.74% of the patients' responses were missing.

Because patients with lower CD4 counts are more likely to drop out from the scheduled study visits (Yuan and Yin (2010)), we believe that the missingness of Y is likely dependent on Y itself. That is, Y is nonignorable missing (Zhang and Wang (2020)). We use the proposed estimator $\hat{\mu}$ and the Morikawa and Kim (2016) estimator $\tilde{\mu}_t$ to estimate the mean of the CD4 cell counts of the patients in Arm I; the estimations for the other arms are similar and omitted.

We take $\mathbf{X} = (X_3, X_4, X_6)$, and choose $f(y|\mathbf{x}, \xi)$ as the normal density, with mean $\mu(\mathbf{x}, \xi) = \xi_1 + \xi_2 x_3 + \xi_3 x_4 + \xi_4 x_6 + \xi_5 x_4^2$ and a constant variance $\sigma(\mathbf{x}, \xi) = \xi_6$. This model is used for all methods shown here, together coupled with the Bayesian information criterion, among the six covariates and their quadratic terms. Because $f(y|\mathbf{x}, \xi)$ is a normal model, checking its correctness can be achieved by testing the normality of the residuals. Three commonly used normality tests, namely, the Shapiro–Wilk test, Kolmogorov–Smirnov test, and Anderson–Darling test, give p-values 0.1422, 0.8547, and 0.2646, respectively, all supporting the pos-

Table 4. Point estimates and interval estimates of the mean of CD496 cell counts (Y) of the patients in Arm I for the ACTG175 data.

	Model I		Model II	
	Point Estimate	Interval Estimate	Point Estimate	Interval Estimate
$\bar{y}_r$	287.62	[269.91, 305.32]	287.62	[269.91, 305.32]
$\hat{\mu}$	258.14	[198.26, 318.02]	256.25	[220.39, 292.11]
$\tilde{\mu}_t$	258.53	[168.55, 348.50]	255.34	[198.82, 311.85]

tulated normal model for $f(y|\mathbf{x}, \xi)$ at the 5% significance level.

We consider two models for the missing probability model:

$$\text{Model I}: \quad \text{pr}(D = 0|\mathbf{x}, y) = \frac{\exp(\alpha^* + x_3\beta_1 + x_4\beta_2 + x_6\beta_3 + y\gamma)}{1 + \exp(\alpha^* + x_3\beta_1 + x_4\beta_2 + x_6\beta_3 + y\gamma)},$$
$$\text{Model II}: \quad \text{pr}(D = 0|\mathbf{x}, y) = \frac{\exp(\alpha^* + x_4\beta_1 + x_6\beta_2 + y\gamma)}{1 + \exp(\alpha^* + x_4\beta_1 + x_6\beta_2 + y\gamma)}.$$

In Model I, there is no instrumental variable, and in Model II, X_3 is the instrumental variable. According to Corollary 1, all model parameters are identifiable.

Because Y is subject to missingness, directly checking the validation of the proposed missing probability model is infeasible. Instead, we achieve this purpose indirectly by checking the validation of the DRM (2.5). The Qin and Zhang (1997) Kolmogorov–Smirnov test produces test statistics 3.03 and 3.01 for the goodness of fit of Models I and II, respectively. Based on 1,000 bootstrap samples, their p-values are found to be 0.335 and 0.397, respectively, which support partially the assumed logistic models for the missing mechanism.

We report the point and Wald interval estimates (at the 95% confidence level) for the mean CD4 cell counts of the patients in Arm I in Table 4. The results of the naive estimator $\bar{y}_r$ are also included. The asymptotic standard deviation of each estimator was estimated based on 1,000 bootstrap samples. We observe that the proposed estimate $\hat{\mu}$ and that of Morikawa and Kim (2016) $\tilde{\mu}_t$ are quite close. However, the proposed interval estimates have much smaller lengths. The naive estimator $\bar{y}_r$ seems to have an upward bias, because we have justified the nonignorable missing mechanism.

Supplementary Material

The online Supplementary Material contains necessary regularity conditions, proofs of Corollary 1 and Theorems 1–3, and additional simulation results.

Acknowledgments

The authors thank the editor, associate editor, and two referees for their constructive comments and suggestions. Dr. Liu's research was supported by the National Natural Science Foundation of China (11771144, 11971300, 11871287), State Key Program of the National Natural Science Foundation of China (719310-04), development fund for Shanghai talents, and 111 project (B14019). Dr. Li's research was supported, in part, by NSERC Grant RGPIN-2020-04964. The first two authors contributed equally to this work.

References

Ai, C., Linton, O. and Zhang, Z. (2018). A simple and efficient estimation method for models with nonignorable missing data. *Statistica Sinica*. In press. DOI: 10.5705/ss.202018.0107.

Anderson, J. A. (1979). Multivariate logistic compounds. *Biometrika* **66**, 17–26.

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

Bondell, H. D. (2007). Testing goodness-of-fit in logistic case-control studies. *Biometrika* **94**, 487–495.

Cai, S., Chen, J. and Zidek, J. V. (2017). Hypothesis testing in the presence of multiple samples under density ratio models. *Statistica Sinica* **27**, 761–783.

Chang, T. and Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika* **95**, 557–571.

Chen, J. and Liu, Y. (2013). Quantile and quantile-function estimations under density ratio model. *Annals of Statistics* **41**, 1669–1692.

Cheng, K. F. and Chu, C. K. (2004). Semiparametric density estimation under a two-sample density ratio model. *Bernoulli* **10**, 583–604.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistician Society, Series B (Methodological)* **30**, 248–275.

Greenlees, J. S., Reece, W. S. and Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association* **77**, 251–261.

Groves, R. M., Presser, S. and Dipko, S. (2004). The role of topic interest in survey participation ecisions. *Public Opinion Quarterly* **68**, 2–31.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundaker, H., Schooley, R. T., Haubrich, R. H., et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *The New England Journal of Medicine* **335**, 1081–1090.

Kenward, M. G. and Molenberghs, G. (1988). Likelihood based frequentist inference when data are missing at random. *Statistica Sinica* **13**, 236–247.

Kim, J. K. and Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association* **106**, 157–165.

Kott, P. S. and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association* **105**, 1265–1275.

Lee, B. and Marsh, L. C. (2000). Sample selection bias correction for missing response observations. *Oxford Bulletin of Economics and Statistics* **62**, 305–322.

Little, R. J. A. (1985). A note about models for selectivity bias. *Econometrica* **53**, 1469–1474.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Inference with Missing Data*. 2nd edition. Wiley, Hoboken, NJ.

Liu, D. and Zhou, X. (2010). A model for adjusting for nonignorable verification bias in estimation of ROC curve and its area with likelihood-based approach. *Biometrics* **66**, 1119–1128.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **44**, 226–233.

Miao, W., Ding, P. and Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association* **111**, 1673–1683.

Morikawa, K. and Kim, J. K. (2016). Semiparametric adaptive estimation with nonignorable nonresponse data. *arXiv preprint arXiv:1612.09207*.

Morikawa, K., Kim, J. K. and Kano, Y. (2017). Semiparametric maximum likelihood estimation with data missing not at random. *The Canadian Journal of Statistics* **45**, 393–409.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *Annals of Statistics* **18**, 90–120.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman and Hall/CRC, New York.

Qin, J., Leung, D. and Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of the American Statistical Association* **97**, 193–200.

Qin, J. and Zhang, B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* **84**, 609–618.

Riddles, M. K., Kim, J. K. and Im, J. (2016). A propensity-score-adjustment method for nonignorable nonresponse. *Journal of Survey Statistics and Methodology* **4**, 215–245.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine* **16**, 285–319.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Shao, J. (2018). Semiparametric propensity weighting for nonignorable nonresponse: A discussion of 'Statistical inference for nonignorable missing data problems: A selective review' by Niansheng Tang and Yuanyuan Ju. *Statistical Theory and Related Fields* **2**, 141–142.

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

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

Tang, N. and Ju, Y. (2018). Statistical inference for nonignorable missing-data problems: A selective review. *Statistical Theory and Related Fields* **2**, 105–133.

Tang, N., Zhao, P. and Zhu, H. (2014). Empirical likelihood for estimating equations with nonignorably missing data. *Statistica Sinica* **24**, 723–747.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York.

Wang, S., Shao, J. and Kim, J. K. (2014). Empirical distributions in selection bias models. *Statistics Sinica* **24**, 1097–116.

Yang, L. (2019). Diagnostics for regression models with discrete outcomes using surrogate empirical residual distribution functions. *arXiv preprint arXiv:1901.04376v2*.

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

Zhang, T. and Wang, L. (2020). Smoothed empirical likelihood inference and variable selection for quantile regression with nonignorable missing response. *Computational Statistics and Data Analysis*. In press.

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

Zhao, H., Zhao, P. and Tang, N. (2013). Empirical likelihood inference for mean functionals with nonignorably missing response data. *Computational Statistics and Data Analysis* **66**, 101–116.

Yukun Liu

KLATASDS - MOE, School of Statistics, East China Normal University, Shanghai 200241, China.

E-mail: ykliu@sfs.ecnu.edu.cn

Pengfei Li

Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, ON N2L 3GI, Canada.

E-mail: pengfei.li@uwaterloo.ca

Jing Qin

National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA.

E-mail: jingqin@niaid.nih.gov

(Received July 2019; accepted June 2020)