Received: 5 April 2022 | Accepted: 3 May 2023

DOI: 10.1111/biom.13881

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Instability of inverse probability weighting methods and a remedy for nonignorable missing data

Pengfei Li[1] | **Jing Qin**[2] | **Yukun Liu**[3]

[1]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada

[2]National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA

[3]KLATASDS – MOE, School of Statistics, East China Normal University, Shanghai, China

Correspondence
Yukun Liu, KLATASDS – MOE, School of Statistics, East China Normal University, Shanghai 200062, China.
Email: ykliu@sfs.ecnu.edu.cn

Funding information
National Key R&D Program of China, Grant/Award Numbers: 2021YFA1000100, 2021YFA1000101; the 111 project, Grant/Award Number: B14019; Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: RGPIN-2020-04964; National Natural Science Foundation of China, Grant/Award Numbers: 12171157, 32030063, 71931004

Abstract

Inverse probability weighting (IPW) methods are commonly used to analyze nonignorable missing data (NIMD) under the assumption of a logistic model for the missingness probability. However, solving IPW equations numerically may involve nonconvergence problems when the sample size is moderate and the missingness probability is high. Moreover, those equations often have multiple roots, and identifying the best root is challenging. Therefore, IPW methods may have low efficiency or even produce biased results. We identify the pitfall in these methods pathologically: they involve the estimation of a moment-generating function (MGF), and such functions are notoriously unstable in general. As a remedy, we model the outcome distribution given the covariates of the completely observed individuals semiparametrically. After forming an induced logistic regression (LR) model for the missingness status of the outcome and covariate, we develop a maximum conditional likelihood method to estimate the underlying parameters. The proposed method circumvents the estimation of an MGF and hence overcomes the instability of IPW methods. Our theoretical and simulation results show that the proposed method outperforms existing competitors greatly. Two real data examples are analyzed to illustrate the advantages of our method. We conclude that if only a parametric LR is assumed but the outcome regression model is left arbitrary, then one has to be cautious in using any of the existing statistical methods in problems involving NIMD.

KEYWORDS
density ratio model, inverse probability weighting, location-scale model, logistic regression, nonignorable missing data

1 | INTRODUCTION

Problems involving nonignorable missing data (NIMD) have attracted much attention in recent years (Tang & Ju, 2018) because they are encountered frequently in many areas but are much more challenging to handle than are problems involving missing at random (MAR) data. Data are MAR if the missingness probability depends only on the observed data and not on the unobserved data; otherwise they are nonignorable missing or missing not at random (Little & Rubin, 2002; Rubin, 1987). We assume that the outcome variable Y may be missing and that the covariate X is always observable. Denote the missingness indicator of Y as R, which equals 0 if Y is missing or 1 otherwise.

Two facts make NIMD more challenging to handle than MAR data. First, for MAR data, the (conditional) missingness probability or propensity score $\text{pr}(R = 1|X, Y) =$

$\text{pr}(R=1|X)$ and the (conditional) outcome regression function $\text{pr}(Y|X)$ are separable in the likelihood, therefore inferences can be made for either of them without the need to know the other; this is not the case for NIMD because $\text{pr}(R=1|X,Y)$ and $\text{pr}(Y|X)$ are entangled together in the likelihood. Second, no identifiability issue is involved in MAR data problems, whereas models may not be identifiable based on NIMD even if fully parametric models are postulated on both $\text{pr}(R=1|X,Y)$ and $\text{pr}(Y|X)$ (Greenless et al., 1982; Heckman, 1979; Miao et al., 2016).

The identification issue must be overcome in NIMD problems before valid statistical inferences are made, otherwise the inference target is vague and meaningless. Robins and Ritov (1997) pointed out that it is impossible to identify the underlying parameters based on NIMD if both $\text{pr}(R=1|X,Y)$ and $\text{pr}(Y|X)$ are left completely unspecified. Attention has been paid to the case in which one of these probabilities is parametric or semiparametric and the other is left unspecified (Chang & Kott, 2008; Kott & Chang, 2010; Kim & Yu, 2011; Morikawa et al., 2017; Morikawa & Kim, 2021; Qin et al., 2002; Riddles et al., 2016; Tang et al., 2003; Wang et al., 2014). When no general identification results are available for NIMD, the joint distribution of the full data can be identified only under specific model assumptions. A popular and general condition for model identifiability with NIMD is the existence of an "instrumental variable" (Wang et al., 2014) or equivalently an "ancillary variable" (Miao & Tchetgen Tchetgen, 2016), which does not affect the missingness probability but may affect the outcome regression function. Miao et al. (2019) found that the full data distribution can be identified with the aid of a shadow variable, which does not affect the missingness probability but may affect the observed outcome regression function $\text{pr}(Y|X,R=1)$.

In recent years, many estimation approaches have been developed for identifiable model parameters based on NIMD (Tang & Ju, 2018; Wang & Kim, 2021). Of those, inverse probability weighting (IPW) methods are the ones used most commonly to deal with missing data including both MAR data (Seaman & White, 2011) and NIMD under a parametric or semiparametric model for the missingness probability. Special cases include augmented IPW estimation or double robust estimation for MAR data (Robins et al., 1995), and the generalized method of moments with an instrumental variable for NIMD (Shao, 2018; Shao & Wang, 2016; Wang et al., 2014; Zhao & Shao, 2015). Double robust estimation under NIMD has also been investigated in the presence of an instrumental or ancillary variable (Ai et al., 2020; Liu et al., 2022; Miao & Tchetgen Tchetgen, 2016; Morikawa & Kim, 2021).

Playing a central role in IPW methods are unbiased IPW estimating equations. However, solving IPW equations numerically may involve nonconvergence problems when the sample size is moderate and the missingness probability is high. In addition, these equations often have multiple roots, and it is challenging to identify the best root as a parameter estimate. Therefore, IPW methods may have low efficiency or even produce biased results. In this paper, under the most popular logistic regression (LR) model for the missingness probability, we find that the instability of IPW methods arises mainly from the fact that IPW equations involve the estimation of a moment-generating function (MGF) (Section 2), and such functions are notoriously unstable in general. As a remedy, we propose modeling the conditional outcome distribution given the covariates of the completely observed individuals by an accelerated time regression model or a semiparametric location-shift model. After transforming the model assumptions to an LR model for the missingness status of the outcome and covariate, we propose estimating the underlying parameters in the propensity score by maximizing a conditional likelihood (Section 3). Our estimation procedure circumvents the estimation of an MGF and hence overcomes the instability of IPW methods. For clarity, all technical details are postponed to the supporting information.

2 | INSTABILITY OF INVERSE PROBABILITY WEIGHTING METHODS

The instability of IPW methods for estimating $E(Y)$ has been well recognized in the literature when the missing mechanism is MAR and $\text{pr}(R=1|X=x)$ is consistently estimated; see Ma and Wang (2020) and references therein. In this section, we begin by discussing the instability of IPW method for estimating the unknown parameters in $\text{pr}(R=1|X=x,Y=y)$ when the missing mechanism is nonignorable.

Suppose that the missingness probability satisfies the commonly used LR model

$$\text{pr}(R=1|x,y)=\text{pr}(R=1|X=x,Y=y)=\frac{1}{1+\exp(\alpha_0+x_1^\top\beta+y\gamma)}, \quad (1)$$

where x_1 is equal to either x or a subvector of x. Suppose that we have n independent and identically distributed random vectors $(r_i,x_i,y_i)\,(1\le i\le n)$ from model (1). In the absence of missing data, the score function of (α_0,β,γ) is

$$-\sum_{i=1}^{n}\{r_i-\text{pr}(R=1|x_i,y_i)\}\times(1,x_{i1}^\top,y_i)^\top.$$

Because $\text{pr}(R=1|x_i,y_i)$ is uniformly bounded, the estimator of (α_0,β,γ) that solves the score equations usually behaves stably.

In the presence of missing response (the y_i with $r_i = 0$ are missing), for any function $g(x)$, the IPW equation

$$\sum_{i=1}^{n} \left\{ \frac{r_i}{\text{pr}(R=1|x_i, y_i)} - 1 \right\} g(x_i) = 0$$

is equivalent to

$$\sum_{i=1}^{n} \left\{ r_i \exp(\alpha_0 + x_{i1}^\top \beta + y_i \gamma) + r_i - 1 \right\} g(x_i) = 0. \quad (2)$$

Essentially, the first term involves estimating the MGF of $\text{pr}(y|x, R=1) = \text{pr}(Y=y|X=x, R=1)$. In the statistical literature, high-order moment estimates are notoriously unstable, let al estimates of an MGF, and it is this unstable estimation that often causes the resulting IPW estimator to perform unstably.

For illustration, we consider a toy example, which is Example 1 in Section 4 with $\alpha_0 = -1.7$ and $\epsilon \sim N(0, 4)$, that is,

$$\begin{aligned} \text{pr}(R=1|x, y) &= \frac{1}{1 + \exp(-1.7 - 0.4x_1 + 0.5y)}, \\ Y|X = x, R = 1 &\sim N(2.5 - x_1 + 1.5x_2, 4), \end{aligned} \quad (3)$$

where $X_1 \sim N(1, 1)$, $X_2 \sim N(0, 1)$, and they are independent. The missingness rate is around 33.9%. An IPW estimator of $(\alpha_0, \beta, \gamma)$ can be obtained by solving

$$\sum_{i=1}^{n} \{ r_i \exp(\alpha_0 + x_{i1}\beta + y_i \gamma) + r_i - 1 \} (1, x_{i1}, x_{i2})^\top = 0.$$

Based on 2000 simulation repetitions for the IPW method with $n = 500$, we have the following observations: (1) the IPW method has multiple roots for 171 times, and does not converge for 74 times and (2) the IPW estimate of γ lies outside $[-3, 3]$ for three times, although its true value is 0.5. If we exclude the above 248 simulated samples, based on the remaining 1752 samples, the relative biases (RBs) of the IPW estimates of α_0, β, and γ are 14.22%, 7.35%, 10.75%, respectively, with standard deviations of 1.14, 0.26, and 0.20, respectively.

The above example shows that the IPW approach by assuming a parametric model for the propensity score function only may not be good enough to estimate the underlying parameters accurately. To overcome this issue, we can make a parametric model assumption, say $f(y|x, \xi)$, on $\text{pr}(y|x, R=1)$ (Kim & Morikawa, 2022; Lee & Marsh, 2000; Liu et al., 2022; Riddles et al., 2016) in addition to the LR model (1). The instability of the IPW method can then be overcome using the method due to Liu et al. (2022). A brief explanation follows, in which because the parameter ξ can be estimated consistently by the maximizer of the conditional likelihood $\prod_{i:r_i=1} f(y_i|x_i, \xi)$ under mild conditions, we take it as known for ease of presentation.

Model (1) implies $\text{pr}(R=0|x, y)/\text{pr}(R=1|x, y) = \exp(\alpha_0 + x_1^\top \beta + y\gamma)$. By Lemma 8.1 of Kim and Shao (2021), we have

$$\begin{aligned} \frac{\text{pr}(R=0|x)}{\text{pr}(R=1|x)} &= E\left\{ \frac{\text{pr}(R=0|X, Y)}{\text{pr}(R=1|X, Y)} \middle| X = x, R = 1 \right\} \\ &= \exp\{\alpha_0 + x_1^\top \beta + c(x; \gamma, \xi)\}, \end{aligned} \quad (4)$$

where $c(x; \gamma, \xi) = \log\{E(e^{\gamma Y}|X = x, R = 1)\}$. Note that (4) implies a new LR model

$$\text{pr}(R=1|x) = \text{pr}(R=1|X=x) = \frac{1}{1 + \exp\{\alpha_0 + x_1^\top \beta + c(x; \gamma, \xi)\}}. \quad (5)$$

The data $\{(x_i, r_i), i = 1, 2, \ldots, n\}$ are all completely observed and follow the new LR model (5). Based on such data, the score equations for $(\alpha_0, \beta, \gamma)$ under model (5) are

$$0 = -\sum_{i=1}^{n} \{ r_i - \text{pr}(R=1|x_i) \} \left(1, x_{i1}^\top, \nabla_\gamma c(x_i; \gamma, \xi)\right)^\top, \quad (6)$$

where $\nabla_\gamma = \partial/\partial\gamma$. Again, because $\text{pr}(R=1|x_i)$ is uniformly bounded, the resulting estimator of $(\alpha_0, \beta, \gamma)$ that solves the score equations behaves more stably than does the IPW estimator.

As we demonstrate in Section 4, Liu et al.'s (2022) method seems to be sensitive to the parametric model assumption on $\text{pr}(y|x, R=1)$. To alleviate this issue, in Section 3, we consider a semiparametric model assumption for $\text{pr}(y|x, R=1)$, and then propose a maximum conditional likelihood method to estimate the unknown parameters in (1) and $E(Y)$. Under the toy example in (3), the newly proposed method did not encounter the multiple roots, nonconvergence, or unstable estimation problems. Further, based on all 2000 repetitions, the RBs of the proposed estimates of α_0, β, γ are 1.91%, 1.21%, and 1.55%, respectively, with standard deviations of 0.42, 0.15, and 0.09, respectively. Clearly, the proposed estimates have much smaller RBs and standard deviations than the IPW estimates.

3 | SEMIPARAMETRIC APPROACH

We assume the same LR model as in (1) for the missingness probability, but we relax the parametric model $f(y|x, \xi)$ for the conditional density function of Y given $X = x$ and $R =$

1 to a semiparametric location-scale model

$$y = \mu(x;\xi) + \epsilon, \tag{7}$$

where $\mu(x;\xi)$ is a known function up to ξ and $E(\epsilon) = 0$. Here, the distribution of ϵ, denoted as $f_\epsilon(\cdot)$, is completely unknown and ϵ is independent of X given $R = 1$. We wish to estimate $\tau = E(Y)$, and model (7) decreases to some extent the risk of model misspecification of a fully parametric model on $p(y|x, R = 1)$.

Throughout this section, we assume that both models (1) and (7) are satisfied, under which we discuss the identifiability issue of the underlying parameters and present our estimation procedure.

3.1 | Parameter identifiability

As we pointed out in the introduction, the issue of parameter identifiability always exists in NIMD problems regardless of which model assumptions are made. Before conducting valid statistical inference about the model parameters in models (1) and (7) and τ, we need to investigate under what conditions they are identifiable.

Because the data with $R = 1$ are all observed, we assume that ξ and f_ϵ in model (7) are identifiable. Without loss of generality, we further regard ξ and f_ϵ as known in this subsection. It remains to study the identifiability of $(\alpha_0, \beta, \gamma)$.

Let $M_1(t) = E(e^{t\epsilon})$ be the MGF of ϵ. Under models (1) and (7), $c(x;\gamma,\xi) = \gamma\mu(x;\xi) + \log M_1(\gamma)$, and the LR model (5) becomes

$$\pi(x;\theta,\xi) = \mathrm{pr}(R = 1|x) = \frac{1}{1 + \exp\{\alpha + x_1^\top\beta + \gamma\mu(x;\xi)\}}, \tag{8}$$

where $\alpha = \alpha_0 + \log\{M_1(\gamma)\}$ and $\theta = (\alpha, \beta^\top, \gamma)^\top$. Because f_ϵ is identifiable, the identifiability of $(\alpha_0, \beta, \gamma)$ is equivalent to that of θ.

Both R and X are observed, $\mathrm{pr}(R = 1|x)$ is identifiable. With (8), θ is identifiable if and only if $\mu(x;\xi)$ is not a linear function of x_1. Let $x = (x_1^\top, x_2^\top)^\top$. If x_2 is not empty, then it is called an instrumental or shadow variable (Miao & Tchetgen Tchetgen, 2016). Here are two special cases: (i) if $\mu(x;\xi)$ is a nonlinear function of x, then θ is identifiable even if there is no instrumental variable or x_2 is empty; (ii) if $\mu(x;\xi)$ is a linear function of x and x_2 is not empty, that is, there exists an instrumental variable, then θ is identifiable.

3.2 | Estimation of model parameters and τ

In this subsection, we propose using a two-step procedure to estimate the model parameters in (1) and (7) based on $\{(y_i r_i, x_i, r_i)\}_{i=1}^n$, where y_i is observed if $r_i = 1$ or missing otherwise.

In step 1, we estimate the unknown parameter ξ in model (7) by the least squares estimator

$$\hat{\xi} = \arg\min_{\xi} \sum_{i=1}^{n} r_i\{y_i - \mu(x_i;\xi)\}^2. \tag{9}$$

This step is implemented easily with the R function `lm` when $\mu(x;\xi)$ is a linear function of x or `nls` when $\mu(x;\xi)$ is a nonlinear function of x.

In step 2, we estimate the unknown parameters $(\alpha_0, \beta, \gamma)$ in model (1). Note that the conditional likelihood of $\{r_i\}_{i=1}^n$ given $\{x_i\}_{i=1}^n$ is

$$l_n(\theta,\xi) = \sum_{i=1}^{n} [r_i \log\{\pi(x;\theta,\xi)\} + (1 - r_i)\log\{1 - \pi(x;\theta,\xi)\}].$$

Instead of directly estimating α_0, we treat $\alpha = \alpha_0 + \log\{M_1(\gamma)\}$ as a new unknown parameter and estimate $\theta = (\alpha, \beta^\top, \gamma)^\top$ first by maximizing the conditional likelihood $l_n(\theta,\xi)$ with ξ replaced by $\hat{\xi}$, that is,

$$\hat{\theta} = \arg\max_{\theta} l_n(\theta, \hat{\xi}). \tag{10}$$

An obvious advantage of this strategy is that we do not need to evaluate $\log\{M_1(\gamma)\}$ in the estimation of α. As $l_n(\theta, \hat{\xi})$ can be regarded as the log-likelihood function under the standard LR model with $\{r_i\}_{i=1}^n$ being the response and $\{x_{i1}, \mu(x_i;\hat{\xi})\}_{i=1}^n$ being the covariates, this step can be implemented with the R function `glm`. Once $(\hat{\theta}, \hat{\xi})$ are obtained, we estimate α_0 by $\hat{\alpha}_0 = \hat{\alpha} - \log\hat{M}_1(\hat{\gamma})$, where $\hat{M}_1(t) = \sum_{i=1}^n r_i e^{t\hat{\epsilon}_i} / \sum_{i=1}^n r_i$ with $\hat{\epsilon}_i = y_i - \mu(x_i;\hat{\xi})$. Kim and Shao (2021) also discussed the conditional likelihood $l_n(\theta,\xi)$. However, they used it not for parameter estimation but only to motivate the identifiability issue with NIMD.

We provide some insights on why the proposed method works and how it remedies the problems associated with the IPW method. In step 1 of the proposed method, as long as ξ is identifiable and $\sum_{i=1}^n r_i$ is not too small, then $\sum_{i=1}^n r_i\{y_i - \mu(x_i;\xi)\}^2$ usually has an unique minimum point and the least square estimator $\hat{\xi}$ is very stable based on our numerical experience. In step 2, $l_n(\theta, \hat{\xi})$ is strictly concave and the maximum point of $l_n(\theta, \hat{\xi})$ is unique, if x_1 and $\mu(x;\hat{\xi})$ are not highly linearly correlated, which is ensured when θ is identifiable and ξ can be estimated reasonably well in step 1. In that situation, the estimator $\hat{\theta}$ is also quite stable. In summary, as long as both θ and ξ are identifiable and $\sum_{i=1}^n r_i$ is not too small, the proposed estimators $(\hat{\theta}, \hat{\xi})$ are uniquely defined and very stable numerically, leading to a stable estimator $\hat{\alpha}_0$ of α_0.

In Section 1 of the supporting information, we analyze three simulated datasets of sample size 500, which are generated from the toy example in (3). In the estimation of $(\alpha_0, \beta, \gamma)$ based on these datasets, the IPW method has either the multiple roots, nonconvergence, or instability problem. In contrast, the proposed method always works stably.

Next, we present the proposed estimator for the target parameter $\tau = E(Y)$, which can be expressed as a function of $\eta = \text{pr}(R = 1)$, ξ, and θ. Under model (7), $E(Y|X = x, R = 1) = \mu(x; \xi)$. Define $M_2(t) = E(\epsilon e^{t\epsilon})$. If model (1) also holds, then

$$E(Y|X = x, R = 0) = \mu(x; \xi) + M_2(\gamma)/M_1(\gamma) := m_0(x; \xi, \gamma); \tag{11}$$

see Section 2 of the supporting information for a proof. Therefore,

$$E(Y|X, R) = R \cdot \mu(X; \xi) + (1 - R) \cdot m_0(X; \xi, \gamma).$$

By the law of total expectation, we have

$$\tau = E\{R \cdot \mu(X; \xi) + (1 - R) \cdot m_0(X; \xi, \gamma)\} = E\{\mu(X, \xi)\} + (1 - \eta)\frac{M_2(\gamma)}{M_1(\gamma)}, \tag{12}$$

where we have used equality (11) and $\eta = E(R)$. Note that (12) suggests that a natural estimators for τ is

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \mu(x_i, \hat{\xi}) + (1 - \hat{\eta})\frac{\hat{M}_2(\hat{\gamma})}{\hat{M}_1(\hat{\gamma})}, \tag{13}$$

where $\hat{\eta} = \sum_{i=1}^{n} r_i/n$ and $\hat{M}_2(t) = \sum_{i=1}^{n} r_i \hat{\epsilon}_i e^{t\hat{\epsilon}_i} / \sum_{i=1}^{n} r_i$.

Remark 1. The proposed estimator $\hat{\tau}$ relies on the regression model (7) and the missingness probability model (1). Based on the completely observed data $\{(y_i r_i, x_i, r_i)\}_{i=1}^{n}$, the correctness of (7) can be verified by using the score test for non-constant error variance proposed by Breusch and Pagan (1979) and Cook and Weisberg (1983); this test is available in R as `ncvTest`. Because we do not have the observed values for Y when $r_i = 0$, model (1) cannot be verified directly; nevertheless, it can be transformed into the LR model (8), and hence we can check (1) indirectly by testing the goodness of fit of the LR model (8). Hosmer et al. (1997) compared several goodness-of-fit tests for the LR model and recommended the one involving the unweighted sum of squares (USS) (le Cessie & van Houwelingen, 1995); this test is available as `resid` in the R package `rms`.

3.3 | Asymptotics

In this subsection, we establish the asymptotic distributions of the proposed estimators $\hat{\xi}$, $\hat{\theta}$, and $\hat{\tau}$. Let η_0, ξ_0, θ_0, and τ_0 be the true values of η, ξ, θ, and τ, respectively, let $B^{\otimes 2} = BB^\top$ for any matrix or vector B, and let $\pi(x) = \pi(x; \theta_0, \xi_0)$. Define $\phi(x; \theta, \xi) = \alpha + x_1^\top \beta + \gamma\mu(x; \xi)$ and $A_1 = E[R\{\nabla_\xi \mu(X; \xi_0)\}^{\otimes 2}]$, $A_2 = E[\pi(X)\{1 - \pi(X)\}\{\nabla_\theta \phi(X; \theta_0, \xi_0)\}^{\otimes 2}]$, $A_3 = E[\pi(X)\{1 - \pi(X)\}\{\nabla_\theta \phi(X; \theta_0, \xi_0)\}\{\nabla_\xi \mu(X; \xi_0)\}^\top]$, and $A_4 = E\{\nabla_\xi \mu(X; \xi_0)\}$.

Theorem 1. *Suppose that models (1) and (7), and the regularity conditions in Section 4 of the supporting information are satisfied. As* $n \to \infty$, $\sqrt{n}(\hat{\xi}^\top - \xi_0^\top, \hat{\theta}^\top - \theta_0^\top)^\top \to N(0, \Sigma)$ *in distribution, where*

$$\Sigma = \begin{pmatrix} \sigma^2 A_1^{-1} & -\gamma_0 \sigma^2 A_1^{-1} A_3^\top A_2^{-1} \\ -\gamma_0 \sigma^2 A_2^{-1} A_3 A_1^{-1} & A_2^{-1} + \gamma_0^2 \sigma^2 A_2^{-1} A_3 A_1^{-1} A_3^\top A_2^{-1} \end{pmatrix}$$

and $\sigma^2 = \mathbb{V}\text{ar}(\epsilon)$.

To use the results in Theorem 1 to construct a Wald-type confidence interval (CI) for the parameters in ξ or θ, we need a consistent estimator for Σ, which can be constructed based on consistent estimators of γ_0, σ^2, and A_1–A_3. Reasonable estimators of γ_0, σ^2, and A_1 are $\hat{\gamma}$, $\hat{\sigma}^2 = \sum_{i=1}^{n} r_i \epsilon_i^2 / \sum_{i=1}^{n} r_i$, and

$$\hat{A}_1 = n^{-1}\sum_{i=1}^{n}\left[r_i\left\{\nabla_\xi \mu(x_i; \hat{\xi})\right\}^{\otimes 2}\right], \tag{14}$$

respectively. The estimators $\hat{A}_2$ and $\hat{A}_3$ for A_2 and A_3 can be constructed in a similar way to (14). Inserting $\hat{\sigma}^2$, $\hat{\xi}$, and $\hat{A}_1$–$\hat{A}_3$ into Σ, we have an estimator $\hat{\Sigma}$ for Σ. With the results in Theorem 1, it can be verified that $\hat{\Sigma}$ is consistent with Σ.

To present the asymptotic distribution of $\hat{\tau}$, we need additional notation. Let $\mu_0 = \mathbb{E}\{\mu(X; \xi_0)\}$, $B_k = \mathbb{E}(R\epsilon^{k-1} e^{\gamma_0 \epsilon})$ for $k = 1, 2, 3$, and $C_k = \mathbb{E}\{R\epsilon^{k-1} e^{\gamma_0 \epsilon} \nabla_\xi \mu(X; \xi_0)\}$ for $k = 1, 2$. Furthermore, let $S = (S_0, S_1^\top, S_2^\top)^\top$ with $S_0 = R - \eta_0$,

$$S_1 = \left(R\epsilon\{\nabla_\xi \mu(X; \xi_0)\}^\top, \{R - \pi(X)\}\{\nabla_\theta \phi(X; \theta_0, \xi_0)\}^\top\right)^\top,$$

$$S_2 = \left(\mu(X; \xi_0) - \mu_0, Re^{\gamma_0 \epsilon} - B_1, R\epsilon e^{\gamma_0 \epsilon} - B_2\right)^\top.$$

It can be verified that $\mathbb{E}(S) = 0$. Denote $V = \mathbb{V}\text{ar}(S) = \mathbb{E}(S^{\otimes 2})$. Finally, let p be the dimension of θ, and let e_p be

a $p \times 1$ vector with the pth element being 1 and the other elements being 0.

Theorem 2. *Suppose that the conditions in Theorem 1 are satisfied. As* $n \to \infty$, $\sqrt{n}(\hat{\tau} - \tau_0) \to N(0, \sigma_\tau^2)$ *in distribution, where* $\sigma_\tau^2 = D^\top V D$. *Here,*

$$D = \left(-\frac{B_2}{B_1}, H_1, H_2, 1, -(1-\eta_0)\frac{B_2}{B_1^2}, \frac{1-\eta_0}{B_1} \right)^\top$$

with

$$H_1 = A_4^\top A_1^{-1} + (1-\eta_0)\left(\frac{B_2}{B_1^2} C_1^\top - \frac{1}{B_1} C_1^\top - \frac{\gamma_0}{B_1} C_2^\top \right) A_1^{-1}$$

$$+ \frac{(1-\eta_0)\gamma_0}{B_1^2}(B_2 - B_1 B_3) e_p^\top A_2^{-1} A_3 A_1^{-1}$$

and $H_2 = (B_2^2 - B_1 B_3) e_p^\top A_2^{-1}(1-\eta_0)/B_1^2$.

To construct a Wald CI for τ based on Theorem 2, we need a consistent estimator of σ_τ^2, which depends on V. We first construct a desirable estimator for V. Let $\hat{\eta} = n^{-1}\sum_{i=1}^n r_i$, $\hat{\mu}_0 = n^{-1}\sum_{i=1}^n \mu(x_i; \hat{\xi})$, $\hat{B}_k = n^{-1}\sum_{i=1}^n (r_i \hat{\epsilon}_i^{k-1} e^{\hat{\gamma}\hat{\epsilon}_i})$, and $\hat{S}_i = (\hat{S}_{0i}, \hat{S}_{1i}^\top, \hat{S}_{2i}^\top)^\top$ with $\hat{S}_{0i} = r_i - \hat{\eta}$,

$$\hat{S}_{1i} = \left(r_i \hat{\epsilon}_i \{\nabla_\xi \mu(x_i; \hat{\xi})\}^\top, \{r_i - \pi(x_i; \hat{\theta}, \hat{\xi})\}\{\nabla_\theta \phi(x_i; \hat{\theta}, \hat{\xi})\}^\top \right)^\top,$$

$$\hat{S}_{2i} = \left(\mu(x_i; \hat{\xi}) - \hat{\mu}_0, r_i e^{\hat{\gamma}\hat{\epsilon}_i} - \hat{B}_1, r_i \hat{\epsilon}_i e^{\hat{\gamma}\hat{\epsilon}_i} - \hat{B}_2 \right)^\top.$$

Then a natural estimator for V is $\hat{V} = n^{-1}\sum_{i=1}^n \hat{S}_i \hat{S}_i^\top$. With the results in Theorem 1, it can be verified that $\hat{V}$ is consistent with V. Using the techniques used to construct $\hat{\Sigma}$, we can construct a consistent estimator $\hat{D}$ for D. Finally, we estimate σ_τ^2 by $\hat{\sigma}_\tau^2 = \hat{D}^\top \hat{V} \hat{D}$, and a $100(1-a)\%$ CI of τ is $\mathcal{I}_\tau = [\hat{\tau} - Z_{1-a/2}\hat{\sigma}_\tau, \hat{\tau} + Z_{1-a/2}\hat{\sigma}_\tau]$, where $Z_{1-a/2}$ is the $(1-a/2)$th quantile of $N(0,1)$.

4 | SIMULATION

4.1 | Setup

We compare the proposed estimator $\hat{\tau}$ of τ with the following competitors developed recently in the literature:

- $\hat{\tau}_P$, the maximum empirical likelihood estimator due to Liu et al. (2022), where the model for Y given $X = x$ and $R = 1$ is (7) with the error distribution being normal;
- $\hat{\tau}_{\text{IPW}}$, the IPW estimator which solves (2) with the form of $g(x)$ being discussed later;
- $\hat{\tau}_{A1}$, the parametric adaptive method due to Morikawa and Kim (2021), where the working model for Y given $X = x$ and $R = 1$ is (7) with the error distribution being normal;
- $\hat{\tau}_{A2}$, the nonparametric adaptive method due to Morikawa and Kim (2021), where the working model for Y given $X = x$ and $R = 1$ is fully nonparametric and the kernel function and bandwidth are those recommended by Morikawa and Kim (2021);
- $\hat{\tau}_{Gk}$, the generalized moment method due to Ai et al. (2020), where the basis functions comprise $\{\prod x_1^{i_1} \cdots x_d^{i_d} : i_1 \geq 0, \ldots, i_d \geq 0, \sum_{j=1}^d i_j \leq k\}$.

We generate data from two examples.

Example 1. Suppose that there are only two covariates $X_1 \sim N(1,1)$ and $X_2 \sim N(0,1)$, which are independent. We set $\text{pr}(R = 1|x, y) = 1/\{1 + \exp(\alpha_0 - 0.4x_1 + 0.5y)\}$ and $y = 2.5 - x_1 + 1.5x_2 + \epsilon$ given $X = x$ and $R = 1$. We consider two values of α_0, that is, -1.7 and -1.2, and the missingness probability increases as α_0 increases.

Example 2. Suppose that there is only one covariate $X \sim N(0,1)$. We set $\text{pr}(R = 1|x, y) = 1/\{1 + \exp(\alpha_0 - 0.4x + 0.5y)\}$ and $y = 2 - x + x^2 + \epsilon$ given $X = x$ and $R = 1$. We consider two values of α_0, that is, -2.7 and -2.2, and again the missingness probability increases as α_0 increases.

In both examples, we consider two distributions for ϵ, that is, $2/3N(-\delta, 4 - 3\delta^2) + 1/3N(2\delta, 4)$ with $\delta = 0$ and 1. The error distribution is just $N(0,4)$ when $\delta = 0$, and it is a normal mixture $2/3N(-1,1) + 1/3N(2,4)$ when $\delta = 1$. The true values of τ and the missingness probability $\text{pr}(R = 0)$ for the two examples are tabulated in Table 1. For the IPW method, we set $g(x)$ in (2) to be $(1, x_1, x_2)^\top$ and $(1, x, x^2)^\top$ in Examples 1 and 2, respectively. The number of repetitions is 2000.

4.2 | Results for point estimates

We summarize the results in terms of RB and mean square error (MSE) for estimating τ in Tables 2 and 3. Note that we encountered numerical problems in implementing the standard IPW method, the adaptive methods due to Morikawa and Kim (2021) and the generalized moment method due to Ai et al. (2020). In the simulation study, we count the number of multiple roots, nonconvergence, or nonreliable cases for each method, where a nonreliable case is one simulation repetition in which either the

TABLE 1 True values of τ and missingness probability pr($R = 0$) in examples 1 and 2.

Example	α_0	δ	τ	$P(R=0)$	Example	α_0	δ	τ	$P(R=0)$
1	−1.7	0	2.177	0.339	1	−1.7	1	2.587	0.369
1	−1.2	0	2.364	0.432	1	−1.2	1	2.868	0.465
2	−2.7	0	3.677	0.338	2	−2.7	1	4.088	0.369
2	−2.2	0	3.869	0.434	2	−2.2	1	4.381	0.469

TABLE 2 Relative bias (RB; ×100), mean square error (MSE; ×100), and number of multiple root, nonconvergence, or nonreliable cases (MNCR) of six estimators of τ (example 1).

	RB	MSE	MNCR	RB	MSE	MNCR	RB	MSE	MNCR	RB	MSE	MNCR
	$\epsilon \sim N(0,4)$											
	$\alpha_0 = -1.7$						$\alpha_0 = -1.2$					
	$n = 500$			$n = 2000$			$n = 500$			$n = 2000$		
$\hat{\tau}$	−0.16	4.18	0	0.17	1.06	0	−0.26	5.64	0	0.04	1.39	0
$\hat{\tau}_P$	0.02	3.95	0	0.18	0.97	0	0.00	5.03	0	0.09	1.23	0
$\hat{\tau}_{\text{IPW}}$	0.11	5.90	248	0.44	1.44	122	0.03	8.12	261	0.52	2.57	121
$\hat{\tau}_{A1}$	−1.12	11.57	13	0.09	1.21	1	−1.77	21.45	14	−0.11	1.72	0
$\hat{\tau}_{A2}$	−30.39	46.39	0	−29.96	43.16	0	−35.88	75.02	0	−35.20	69.99	0
$\hat{\tau}_{G1}$	1.70	7.85	13	0.95	2.28	4	2.00	11.04	16	1.24	4.55	5
$\hat{\tau}_{G2}$	−6.08	6.22	6	−2.15	1.45	0	−8.02	9.74	10	−2.93	2.21	0
	$\epsilon \sim 2/3N(-1,1) + 1/3N(2,4)$											
	$\alpha_0 = -1.7$						$\alpha_0 = -1.2$					
	$n = 500$			$n = 2000$			$n = 500$			$n = 2000$		
$\hat{\tau}$	−0.27	10.09	0	−0.10	2.38	0	−0.67	14.49	0	−0.26	3.60	0
$\hat{\tau}_P$	−13.18	17.17	0	−13.39	13.31	0	−15.20	26.54	0	−15.27	20.96	0
$\hat{\tau}_{\text{IPW}}$	0.11	15.42	428	1.09	5.95	212	-0.06	24.63	475	1.47	10.16	199
$\hat{\tau}_{A1}$	−0.55	32.81	31	−0.22	7.89	4	−1.29	44.22	41	−0.37	7.55	7
$\hat{\tau}_{A2}$	−41.59	118.50	0	−41.37	115.20	0	−47.39	188.29	0	−46.95	182.17	0
$\hat{\tau}_{G1}$	2.39	18.50	49	2.62	9.53	14	2.46	29.47	57	2.97	14.20	13
$\hat{\tau}_{G2}$	−11.83	19.34	12	−5.04	5.17	0	−14.93	33.29	13	−6.47	8.62	0

estimate of τ is outside the range $[-10, 10]$ or the estimate of γ is outside the range $[-3, 3]$. For each method, the RB and MSE results reported are evaluated based on the cases, in which there are no multiple roots, nonconvergence, or nonreliable estimates.

From the simulation results, we make the following observations. (1) The proposed estimator $\hat{\tau}$ performs similarly to the estimator $\hat{\tau}_P$ due to Liu et al. (2022) when the error distribution is normal, while $\hat{\tau}$ has smaller RB and MSE than those of $\hat{\tau}_P$ when the error distribution is a mixture of normal distributions. This shows the robustness of the proposed method. (2) As we discussed before, the IPW method and methods due to Morikawa and Kim (2021) and Ai et al. (2020) may experience numerical issues, which become more prominent when the sample size is small, the missingness probability is high, and/or no instrumental variable exists. (3) Even after the cases producing multiple roots, nonconvergence, or nonreliable estimates are excluded, the IPW estimator $\hat{\tau}_{\text{IPW}}$ still has much larger MSE than our estimator $\hat{\tau}$. 4) The parametric adaptive method due to Morikawa and Kim (2021) produces much larger MSE than does our estimator $\hat{\tau}$, although the former far outperforms their nonparametric adaptive method in terms of both RB and MSE. 5) The performance of the generalized moment method due to Ai et al. (2020) depends on the choice of basis functions. When the number of basis functions increases, the bias increases and the variance decreases. Compared with the proposed method, the generalized moment method usually produces an MSE that is at least 50% and can be double or ever higher.

We also compare the proposed method with the maximum empirical likelihood estimator of Liu et al. (2022) for a binary outcome; see Section 3 of the Supporting Information. In this scenario, model (7) is misspecified for the proposed method, while the models are all correctly specified for Liu et al.'s (2022) method. We observe that

TABLE 3 Relative bias (RB; ×100), mean square error (MSE; ×100), and number of multiple-root, nonconvergence, or nonreliable cases (MNCR) of seven estimators of τ (example 2).

	RB	MSE	MNCR	RB	MSE	MNCR	RB	MSE	MNCR	RB	MSE	MNCR
	$\epsilon \sim N(0,4)$											
	$\alpha_0 = -2.7$						$\alpha_0 = -2.2$					
	$n = 500$			$n = 2000$			$n = 500$			$n = 2000$		
$\hat{\tau}$	0.25	4.99	0	0.14	1.15	0	0.36	7.14	0	0.21	1.68	0
$\hat{\tau}_P$	0.39	4.59	0	0.20	1.06	0	0.55	6.60	0	0.27	1.52	0
$\hat{\tau}_{\text{IPW}}$	3.29	12.29	468	2.2	4.24	160	4.14	19.49	541	3.37	8.69	170
$\hat{\tau}_{A1}$	−1.01	61.34	133	−0.03	2.64	4	−0.87	107.69	195	0.06	7.05	10
$\hat{\tau}_{A2}$	−11.81	22.72	16	−8.71	11.30	3	−14.77	38.04	18	−10.73	18.76	2
$\hat{\tau}_{G2}$	4.25	14.48	79	2.83	5.65	3	5.33	23.20	101	4.17	11.44	1
$\hat{\tau}_{G3}$	−3.99	15.57	144	−1.48	2.84	10	−4.06	20.79	135	−1.54	3.96	16
$\hat{\tau}_{G4}$	−6.96	17.07	92	−3.93	4.38	16	−7.58	24.74	59	−4.61	6.69	5
	$\epsilon \sim 2/3N(-1,1) + 1/3N(2,4)$											
	$\alpha_0 = -2.7$						$\alpha_0 = -2.2$					
	$n = 500$			$n = 2000$			$n = 500$			$n = 2000$		
$\hat{\tau}$	−0.27	11.94	0	0.15	3.05	0	−0.10	19.15	0	0.20	5.18	0
$\hat{\tau}_P$	−8.32	17.87	0	−8.35	13.16	0	−9.68	27.27	0	−9.89	21.00	0
$\hat{\tau}_{\text{IPW}}$	3.38	26.8	863	4.83	18.11	343	3.49	42.34	983	6.06	33.65	466
$\hat{\tau}_{A1}$	−0.70	180.78	208	0.24	21.22	29	0.98	428.26	575	0.76	39.74	52
$\hat{\tau}_{A2}$	−20.57	75.86	100	−16.35	46.33	32	−24.77	124.82	209	−19.90	78.37	32
$\hat{\tau}_{G2}$	4.18	26.72	164	6.28	22.98	16	3.89	42.10	213	8.00	41.48	38
$\hat{\tau}_{G3}$	−6.15	27.81	149	−1.96	8.77	22	−7.40	44.09	139	−2.34	17.44	40
$\hat{\tau}_{G4}$	−9.22	30.05	61	−5.33	12.07	12	−11.37	52.25	41	−6.60	23.75	16

our method has very similar performance to that of Liu et al.'s (2022) method, which implies that our method has certain robustness.

4.3 | Results for confidence intervals

In this subsection, we evaluate the performance of the proposed CI $\mathcal{I}_\tau$ for τ. To improve the performance of $\mathcal{I}_\tau$, we also consider the bootstrap t-type CI, which is $\mathcal{I}_\tau$ with the normal quantile replaced by the corresponding nonparametric bootstrap quantiles based on 1000 bootstrap samples; we denote the bootstrap t-type CI by $\mathcal{I}_\tau^B$. The simulated coverage probabilities of $\mathcal{I}_\tau$ and $\mathcal{I}_\tau^B$ are provided in Table 4. We did not compare the proposed CIs with those based on other methods because $\hat{\tau}_P$ produces biased results when the error distribution is not normal, and the IPW method and the methods due to Morikawa and Kim (2021) and Ai et al. (2020) may experience numerical issues.

Table 4 shows that $\mathcal{I}_\tau$ has accurate coverage probabilities when the error distribution is either normal or nonnormal with $n = 2000$, but it experiences undercoverage when the error distribution is nonnormal with $n = 500$. By contrast, the bootstrap t-type CI has accurate coverage probabilities in all situations and so is recommended in applications.

TABLE 4 Coverage probabilities of $\mathcal{I}_\tau$ and $\mathcal{I}_\tau^B$ at 95% nominal level.

Example	α_0	$\mathcal{I}_\tau$	$\mathcal{I}_\tau^B$	$\mathcal{I}_\tau$	$\mathcal{I}_\tau^B$
	$\epsilon \sim N(0,4)$				
		$n = 500$		$n = 2000$	
1	−1.7	95.0	94.7	95.4	95.0
1	−1.2	94.1	94.8	95.4	94.7
2	−2.7	95.0	95.0	95.5	94.7
2	−2.2	94.8	95.4	95.6	94.8
	$\epsilon \sim 2/3N(-1,1) + 1/3N(2,4)$				
		$n = 500$		$n = 2000$	
1	−1.7	93.2	94.4	95.0	95.2
1	−1.2	92.6	94.4	94.7	95.0
2	−2.7	92.7	94.7	95.0	95.0
2	−2.2	92.6	95.0	94.7	95.4

5 | REAL EXAMPLES

We analyze two real examples for illustration. The first example involves human immunodeficiency virus (HIV) data from the AIDS Clinical Trials Group Protocol 175 (ACTG175) (Hammer et al., 1996), in which $n = 2139$ HIV-infected patients were enrolled. The patients were divided

randomly into four arms according to the regimen of treatment that they received: (I) zidovudine monotherapy, (II) zidovudine + didanosine, (III) zidovudine + zalcitabine, and (IV) didanosine monotherapy. For illustration, we consider the patients in arm III; analysis for the other arms can be conducted similarly.

The data record many measurements from each patient, including their 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 determining the mean of the CD4 cell counts after the patients were treated for about 96 weeks. We take cd496 as the response variable Y, and we take age, weight, cd40, cd420, cd80, and cd820 as covariates $X_1, \ldots, X_6$, respectively. Because of either the trial ending or lack of follow-up, 35.7% of the patients' responses were missing.

We take $X = (X_1, X_3, X_4, X_6)$ and consider the following location-scale model for Y given $X = x$ and $R = 1$:

$$y = \mu(x, \xi) + \epsilon = \xi_1 + \xi_2 x_1 + \xi_3 x_3 + \xi_4 x_4 + \xi_5 x_6 + + \xi_6 x_1^2 + \xi_7 x_4^2 + \epsilon. \quad (15)$$

This model is chosen by the all-subset selection method coupled with the Akaike information criterion among the six covariates and their quadratic terms. The score test for nonconstant error variance proposed by Breusch and Pagan (1979) and Cook and Weisberg (1983) gives a p-value of around 0.560, which supports the assumption of constant error variance. To check whether the normality assumption is suitable for the errors in (15), we perform a Shapiro–Wilk test for the residuals, which produces a p-value of around 4.28×10^{-8}. Therefore, we have no evidence against the location-scale mode with a constant error in (15), but the normal error assumption may not be suitable.

Next, we consider modeling the missingness probability. Recall that the location-scale model (7) and the missingness probability model (1) imply the LR model (8) for R given $X = x$. With the available data $\{(x_i, r_i)\}_{i=1}^n$ and the chosen model $\mu(x; \xi)$, we use the all-subset section method coupled with the Akaike information criterion to choose the most appropriate model for R given $X = x$, which is

$$\mathrm{pr}(R = 1|x) = \frac{1}{1 + \exp\{\alpha_0 + x_1\beta_1 + x_3\beta_2 + \gamma\mu(x; \xi)\}}. \quad (16)$$

To evaluate the goodness of fit of this model, the USS test due to le Cessie and van Houwelingen (1995) gives a p-value of around 0.722. Therefore, model (16) provides a reasonable fit for R given $X = x$, which together with the location-scale model (15) for Y given $X = x$ and $R = 1$ implies that the missingness probability model

$$\mathrm{pr}(D = 1|x, y) = \frac{1}{1 + \exp(\alpha_0 + x_1\beta_1 + x_3\beta_2 + y\gamma)} \quad (17)$$

is suitable.

We now apply the proposed estimator $\hat{\tau}$ and CI $\mathcal{I}_\tau^B$, based on models (15) and (17), to the ACTG175 data for patients in arm III. For comparison, we include $\hat{\tau}_P$, $\hat{\tau}_{\mathrm{IPW}}$, $\hat{\tau}_{A1}$, and $\hat{\tau}_{G1}$ and their corresponding bootstrap percentile CIs based on 1000 bootstrap samples. The IPW method solves (2) with $g(x) = (1, x_1, x_3, \mu(x; \hat{\xi}))^\top$. We chose not to include the nonparametric adaptive method due to Morikawa and Kim (2021) and $\hat{\tau}_{Gk}$ due to Ai et al. (2020) for $k \geq 2$, as both methods show large biases in our simulation study. The results are summarized in Table 5. In the analysis of ACTG175 data, we do not encounter numerical problems in implementing $\hat{\tau}_{\mathrm{IPW}}$. This may explain why the proposed method and the IPW method give similar point estimates. The other three point estimates $\hat{\tau}_P$, $\hat{\tau}_{A1}$, and $\hat{\tau}_{G1}$ are slightly different from the proposed estimate $\hat{\tau}$. Based on our simulation results for the nonnormal case, $\hat{\tau}$ always has quite small biases, so we reason that the result for $\hat{\tau}$ is more reliable. The CI based on $\hat{\tau}$ is the smallest length among the five methods, which shows the advantage of the proposed method for the nonnormal error case.

The second example involves the Peabody Picture Vocabulary Test (PPVT) data analyzed by Chen et al. (2022), which were collected as part of the National Longitudinal Survey of Youth (NLSY79 Child and Young Adult cohort). The PPVT comprises a number of items, each of which involves four pictures; the interviewer says a word out loud, and the child selects the picture of the four that best describes the word's meaning. The data come from test results between 1986 and 1992 for children who were aged between 3 and 4 years at the 1986 assessment and whose mothers reported nonzero income in at least 1 year between 1986 and 1992. In total, $n = 557$ children are in the sample.

We let the response Y be the logarithm of the difference in PPVT score between 1986 and 1992. Following Chen et al. (2022), we consider seven covariates: gender (1 = male, 0 = female; x_1), race (1 = White, 0 = Other; x_2), mother's hourly income (x_3), mother's education (1 $\geq$ 12 years, 0 $\leq$ 12 years; x_4), and three dummy variables (x_5–x_7) that classify the data by the four quartiles of the 1986 PPVT score,

TABLE 5 Analysis results for ACTG175 data (Y = CD496 cell counts of patients in arm III) and PPVT data (Y = logarithm of difference in PPVT score between 1986 and 1992 of children).

	Point estimate	Interval estimate	Length of CI
ACTG175 data			
$\hat{\tau}$	308.98	[279.68, 330.97]	51.29
$\hat{\tau}_P$	305.72	[278.51, 333.44]	54.93
$\hat{\tau}_{\text{IPW}}$	309.00	[276.39, 336.25]	59.86
$\tilde{\tau}_{A1}$	310.39	[281.94, 339.93]	58.00
$\tilde{\tau}_{G1}$	313.68	[285.19, 338.48]	53.29
PPVT data			
$\hat{\tau}$	4.302	[4.280, 4.323]	0.043
$\hat{\tau}_P$	4.303	[4.272, 4.330]	0.059
$\hat{\tau}_{\text{IPW}}$	4.300	[4.266, 4.325]	0.058
$\hat{\tau}_{A1}$	4.301	[4.275, 4.324]	0.049
$\hat{\tau}_{G1}$	4.297	[4.261, 4.327]	0.067

that is, $(x_5, x_6, x_7) = (0,0,0)$, (1,0,0), (0,1,0), or (0,0,1) if a 1986 PPVT score is in the first, second, third, or fourth quartile, respectively. For various reasons, such as motivation, family influence, and perceived poor performance, only 387 valid assessments were obtained in 1992, which gives a missing-data rate of 30.5%.

We take $X = (X_1, \ldots, X_7)$ and consider the following location-scale model for Y given $X = x$ and $R = 1$:

$$y = \mu(x, \xi) + \epsilon = \xi_1 + \xi_2 x_1 + \cdots + \xi_8 x_7 + \epsilon. \quad (18)$$

The p-value of the score test for nonconstant error variance is around 0.296, and that for the Shapiro–Wilk test on residuals is around 5.74×10^{-12}. These results indicate that the location-scale model (18) is reasonable for Y given $X = x$ and $R = 1$ but that the normality assumption for the error may not be suitable.

Following Chen et al. (2022), we use x_5–x_7 as instrumental variables and consider the following missingness probability model:

$$\text{pr}(R = 1|x, y) = \frac{1}{1 + \exp(\alpha_0 + x_1\beta_1 + \cdots + \beta_4 x_4 + y\gamma)}. \quad (19)$$

The USS test due to le Cessie and van Houwelingen (1995) for the goodness of fit of the induced LR model

$$\text{pr}(R = 1|x) = \frac{1}{1 + \exp\{\alpha + x_1\beta_1 + \cdots + \beta_4 x_4 + \gamma\mu(x; \xi)\}} \quad (20)$$

gives a p-value of around 0.737, which supports the missingness probability model (19).

We now apply the proposed estimator $\hat{\tau}$ and CI $\mathcal{I}_\tau^B$, based on (18) and (19), to the PPVT data. For comparison, we also include the results for $\hat{\tau}_P$, $\hat{\tau}_{\text{IPW}}$, $\hat{\tau}_{A1}$, and $\hat{\tau}_{G1}$. The results are summarized in Table 5. It is worth mentioning that we do not encounter numerical problems while implementing $\hat{\tau}_{\text{IPW}}$ in the analysis of PPVT data. This observation may explain the similarity between the proposed estimate and the IPW estimate. The other three methods produced point estimates similar to the proposed method for this data. However, the proposed CI $\mathcal{I}_\tau^B$ has the shortest length among all five methods, which again shows the advantage of the proposed method for non-normal data.

6 | CONCLUDING REMARKS

As mentioned in the beginning of the introduction, inference for problems involving NIMD is much harder than that for those involving MAR data. In general, one must either specify the propensity score or make a parametric assumption about the outcome given covariates, otherwise the underlying models are not identifiable. The existing literature shows that it is possible to identify the propensity-score parameters if one specifies a parametric model for the propensity score only and leaves the regression model arbitrary. However, the IPW estimating equations involve an MGF, and such functions can suffer from slow convergence; consequently, such equations are very unstable numerically and may have multiple roots. Our extensive simulation studies have shown that inference based on a parametric logistic propensity-score model alone is very unreliable, and one must make some assumption about the regression model, either parametrically or semiparametrically. In this paper, we have proposed an innovative method for NIMD problems when a parametric model for the propensity score is specified and the observed outcome follows a location-shift model with an unspecified error distribution. Extensive simulations have shown that our method far outperforms the existing ones.

We end up this paper with some discussion on possible extensions of the proposed method. Kim and Yu (2011) and Kim and Morikawa (2022) considered a semiparametric model for the missingness probability:

$$\text{pr}(R = 1|x, y) = \frac{1}{1 + \exp\{g(x_1) + \gamma y\}}, \tag{21}$$

where $g(\cdot)$ is a completely unspecified function. We may generalize our method to the setup when both (21) and (7) are satisfied. Under these two model assumptions, similar to (8), we have a new partially linear LR model

$$\text{pr}(R = 1|x) = \frac{1}{1 + \exp\{g^*(x_1) + \gamma\mu(x; \xi)\}},$$

where $g^*(x_1) = g(x_1) + \log\{M_1(\gamma)\}$. Our method can be extended using either kernel method, local likelihood method, or spline method to estimate $g(\cdot)$ and γ. Wang et al. (2021) proposed a data-adaptive method to choose instrumental variables instead of assuming they are known in advance. Our method may be extended for this purpose as well. We leave both for future research.

ACKNOWLEDGMENTS

The authors thank Dr. D. H. Y. Leung for sharing the PPVT data and Dr. J. K. Kim for sharing the R code. They also thank the editor, associate editor, and two referees for constructive comments and suggestions that led to significant improvements in the paper. This research is supported by the National Key R&D Program of China (2021YFA1000100 and 2021YFA1000101), the Natural Sciences and Engineering Research Council of Canada (RGPIN-2020-04964), the National Natural Science Foundation of China (71931004, 32030063 and 12171157), and the 111 project (B14019). Dr. Liu is the corresponding author.

DATA AVAILABILITY STATEMENT

The first dataset that supports the findings in this paper is openly available from the `R` package `speff2trial` (https://cran.r-project.org/web/packages/speff2trial/index.html). The second dataset is available from Dr. Denis Heng-Yan Leung, denisleung@smu.edu.sg, the corresponding author of Chen et al. (2022).

ORCID

Pengfei Li https://orcid.org/0000-0003-2165-9157
Jing Qin https://orcid.org/0000-0003-2817-6326
Yukun Liu https://orcid.org/0000-0002-9743-9276

REFERENCES

Ai, C., Linton, O. & Zhang, Z. (2020) A simple and efficient estimation method for models with non-ignorable missing data. *Statistica Sinica*, 30, 1949–1970.

Breusch, T.S. & Pagan, A.R. (1979) A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47, 1287–1294.

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

Chen, X., Leung, D.H.Y. & Qin, J. (2022) Nonignorable missing data, single index propensity score and profile synthetic distribution function. *Journal of Business & Economic Statistics*, 40, 705–717.

Cook, R.D. & Weisberg, S. (1983) Diagnostics for heteroscedasticity in regression. *Biometrika*, 70, 1–10.

Greenless, 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.

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

Heckman, J.J. (1979) Sample selection bias as a specification error. *Econometrica*, 47, 153–161.

Hosmer, D.W., Hosmer, T., Le Cessie, S. & Lemeshow, S. (1997) A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16, 965–980.

Kim, J.K. & Morikawa, K. (2022) *An empirical likelihood approach to reduce selection bias in voluntary samples*. Available at: https://arxiv.org/abs/2211.02998

Kim, J.K. & Shao, J. (2021) *Statistical methods for handling incomplete data*, 2nd edition. Boca Raton, FL: CRC Press.

Kim, J.K. & 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. & Chang, T. (2010) Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105, 1265–1275.

le Cessie, S. & van Houwelingen, J.C. (1995) Testing the fit of a regression model via score tests in random effects models. *Biometrics*, 51, 600–614.

Lee, B. & 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. & Rubin, D.B. (2002) *Statistical inference with missing data*, 2nd edition. Hoboken, NJ: Wiley.

Liu, Y., Li, P. & Qin, J. (2022) Full-semiparametric-likelihood-based inference for non-ignorable missing data. *Statistica Sinica*, 32, 271–292.

Ma, X. & Wang, J. (2022) Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, 115, 1851–1860.

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

Miao, W., Liu, L., Tchetgen Tchetgen, E. & Geng, Z. (2019) *Identification, doubly robust estimation, and semiparametric efficiency theory of nonigorable missing data with a shadow variable*. Available at: https://arxiv.org/abs/1509.02556

Miao, W. & Tchetgen Tchetgen, E. (2016) On varieties of doubly robust estimators under missing not at random with an ancillary variable. *Biometrika*, 103, 475–482.

Morikawa, K. & Kim, J.K. (2021) Semiparametric optimal estimation with nonignorable nonresponse data. *Annals of Statistics*, 49, 2991–3014.

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

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

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

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

Robins, J.M., Rotnitzky, A. & Zhao, L.P. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106–121.

Rubin, D.B. (1987) *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Seaman, S.R. & White, I.R. (2011) Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22, 278–295.

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

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

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

Wang, H. & Kim, J.K. (2021) *Statistical inference after kernel ridge regression imputation under item nonresponse*. Available at: https://arxiv.org/abs/2102.00058

Wang, L., Shao, J. & Fang, F. (2021) Propensity model selection with nonignorable nonresponse and instrument variable. *Statistica Sinica*, 31, 647–672.

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

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

SUPPORTING INFORMATION

The supporting information contains analyses of three simulated datasets, a detailed derivation of equality (11) in Section 3, additional simulation results, regularity conditions, and proofs of Theorems 1 and 2. The three simulated datasets and R code are also available with this paper at the Biometrics website on Wiley Online Library.

Data S1

How to cite this article: Li, P., Qin, J. & Liu, Y. (2023) Instability of inverse probability weighting methods and a remedy for nonignorable missing data. *Biometrics*, 1–12. https://doi.org/10.1111/biom.13881