Statistica Sinica Preprint No: SS-2024-0204	
Title	Identification and Efficient Estimation in Regression Analysis with Response Missing Not At Random
Manuscript ID	SS-2024-0204
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0204
Complete List of Authors	Qinglong Tian, Donglin Zeng and Jiwei Zhao
Corresponding Authors	Jiwei Zhao
E-mails	jiwei.zhao@wisc.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

Identification and Efficient Estimation in Regression Analysis with Response Missing Not At Random

Qinglong Tian[1], Donglin Zeng[2], Jiwei Zhao[3]

[1] *University of Waterloo,* [2] *University of Michigan*

[3] *University of Wisconsin-Madison*

Abstract: Missing-data is a pervasive problem in regression analysis, compromising the accuracy and efficiency of parameter estimates. This paper focuses on the challenging scenario of missing not at random (MNAR) data, where the missingness of a value is linked to the value itself. Traditional approaches to addressing MNAR data confront a trade-off: imposing stringent assumptions about the missingness mechanism can enhance efficiency but curtail robustness, whereas accommodating model misspecification can bolster robustness but at the expense of efficiency. In addition, assuming a nonparametric MNAR mechanism will lead to model identifiability issues. We propose a novel approach that overcomes this limitation. Firstly, we address the model identifiability issue using the shadow variable. Then, by leveraging the sieve method, we can model the MNAR mechanism nonparametrically. This approach achieves the best of both worlds: it gains robustness by avoiding strict assumptions about the missingness mechanism while simultaneously achieving the semiparametric efficiency bound for the parameter of interest (meaning our estimator has the lowest possible

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

asymptotic variance). The paper delves into the theoretical framework, outlining conditions for identifiability, constructing the semiparametric likelihood function, and rigorously proving the estimator's semiparametric efficiency. Additionally, we present an EM-type algorithm for practical implementation, discussing the E-step and M-step iterations and variance estimation methods. Finally, simulations and a real-data application demonstrate the effectiveness of our proposed method compared to existing approaches.

Key words and phrases: Missing Data; Missing Not at Random; Identification; Semiparametric Efficiency; Efficient Estimation; Method of Sieves.

1. Introduction

Regression analysis is a cornerstone of statistical research, revealing relationships between a response variable and one or more covariates. However, real-world data often suffers from missing values. Imagine a sociological survey where participants skip questions or a healthcare study where patients miss appointments. These missing data points can significantly impact regression analysis. Recognizing this challenge, statisticians have developed a robust framework for understanding missing data patterns and their influence. This framework categorizes missingness by the missing mechanisms and offers a suite of methods tailored to different scenarios.

Statisticians use the concept of missing data mechanism to describe

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

how data are missing. In the case of the missing at random (MAR) mechanism, the probability of a data point being missing depends solely on the observed variables in the dataset, not on the missing values themselves. Consider a survey where men are more likely to answer an income question than women. Here, missing income data are considered MAR because the probability of missing data is connected to gender (an observed variable) but not a person's actual income (the missing variable). A nice property of the MAR assumption is that it allows us to estimate regression coefficients using only complete cases (observations with no missing data). However, this approach has a limitation: it can be inefficient, meaning we might not extract all the valuable information from the data (by discarding the incomplete data). To overcome the limitations of complete-case analysis under MAR, statisticians have developed several methods to improve estimation efficiency. These methods include imputation, inverse probability weighting, empirical likelihood (Wang and Chen 2009), and auxiliary variable (Chen et al. 2008). More comprehensive references include Little and Rubin (2002) and Tsiatis (2006).

The MAR assumption, however, has limitations. Consider a survey where people with extremely high or low incomes are less likely to report their income. In this scenario, missing income data would depend on the

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

very value we are missing (income level). This situation exemplifies missing not-at-random (MNAR) data. Here, the probability of a variable being missing hinges on the value of the variable itself, which may not be observed. This makes MNAR data considerably more challenging to analyze. Even if we can accurately model the missingness mechanism (e.g., using a logistic regression model), the estimation process remains complicated because some of the variable's values are absent. Consequently, analyzing MNAR data requires more advanced techniques and stronger assumptions than the MAR case.

Despite of challenges of MNAR mechanisms, researchers have proposed various methods. For example, Kim and Yu (2011) and Tang et al. (2014) considered an exponential tilting model under the MNAR data. Other researchers have developed robust and efficient methods for semiparametric MNAR models, such as the work by Li et al. (2022), Shetty et al. (2023), and Shetty et al. (2025). In the literature, a line of research attempted to model the MNAR mechanism nonparametrically. A seminal work is Tang et al. (2003), whose proposed procedure took advantage of some structural properties among the variables and could avoid estimating the MNAR mechanism model. Zhao and Shao (2015) and Zhao and Shao (2017) generalized the idea in Tang et al. (2003) to the shadow variable framework

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

and proposed a pseudo-likelihood-based method for estimation, and their procedures do not need to estimate the MNAR mechanism. The property that a proposed method does not need an estimate of a nuisance model has some intrinsic relations with the robustness property in semiparametric statistics. Indeed, Zhao and Ma (2022) characterized the semiparametric structure of the model studied in Zhao and Shao (2015) and proposed a versatile estimating procedure that only requires a working model, which can be arbitrarily misspecified, of the MNAR mechanism. Zhao and Ma (2022) also pointed out that the method proposed in Zhao and Shao (2015) is actually their special case when one simplistically chooses the MNAR mechanism working model as a constant. All the work mentioned above is under the umbrella of methods that do not require the correct specification or avoid estimating the MNAR mechanism model. While enjoying the robustness of model misspecification, one generally cannot achieve estimation efficiency under this line of work.

Another type is to model the MNAR mechanism nonparametrically, using a nonparametric method to approximate the MNAR mechanism model. It will generally lead to the semiparametrically efficient estimate of the main parameter of interest without imposing any parametric assumption on the missing data mechanism. However, there is a lack of existing work

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

to formalize the estimation procedure and establish theoretical results on estimation efficiency. In this paper, we will fill this gap by modeling the MNAR mechanism via the sieve method and will study the semiparametrically efficient regression coefficient estimation. It is worthwhile to note that Zhao and Ma (2018) resolved a conjecture regarding some optimality properties (i.e., estimation efficiency) in the estimation procedure of Tang et al. (2003). However, their estimation procedures are still under the umbrella of without estimating the MNAR mechanism model. Thus, their optimality still does not lead to semiparametrically efficient estimation.

Lastly, we briefly highlight the contributions of this paper. In Section 2, recognizing the notorious issue of model identifiability with MNAR data, we first provide a novel condition such that both the main parameter of interest and the nuisance MNAR mechanism model are identifiable. Then, we derive the semiparametric likelihood function, where the missingness mechanism model is approximated using the sieves method. In Section 3, we rigorously prove that the resulting estimator achieves the semiparametric efficiency bound and thus is semiparametrically efficient. In Section 4, we devise an EM-type algorithm for the numerical implementation, and we further provide the details of both E-step and M-step iterations and the variance estimation method. Sections 5 and 6 are the simulation stud-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

ies and real data applications, aiming to illustrate the proposed method's finite-sample performance and its comparison with some existing methods in the literature. All the technical proofs are deferred in the supplementary material.

2. Methodology

Throughout the paper, we consider the regression model $f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})$ where Y is a scalar response, $\mathbf{X}$ is a covariate vector, and $\boldsymbol{\beta}$ is the d-dimensional parameter of interest. In our framework, we consider the situation that $\mathbf{X}$ is fully observed, but Y has missing values. We use R to denote whether Y is observed, i.e., $R = 1$ if Y is observed and $R = 0$ if otherwise. The missingness mechanism model, also called the propensity score model, is $\text{pr}(R = 1 \mid y, \mathbf{x})$. If the propensity score model does not depend on y, the mechanism is called ignorable or missing at random (MAR); that is, $\text{pr}(R = 1 \mid y, \mathbf{x}) = \text{pr}(R = 1 \mid \mathbf{x})$. Otherwise, if $\text{pr}(R = 1 \mid y, \mathbf{x}) \neq \text{pr}(R = 1 \mid \mathbf{x})$, it is called nonignorable or missing not at random (MNAR). This paper focuses on the MNAR missingness, under which one should note that

$$p(Y \mid \mathbf{X}, R = 1) = \frac{\text{pr}(R = 1 \mid Y, \mathbf{X})}{\text{pr}(R = 1 \mid \mathbf{X})} p(Y \mid \mathbf{X}) \neq p(Y \mid \mathbf{X});$$

therefore, the naive regression analysis based on the completely observed subjects (i.e., the left side of the equation above) would generally produce

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

a biased estimate for the regression parameter $\boldsymbol{\beta}$ (i.e., the right side of the equation).

2.1 Model Identifiability

Model identifiability is a notorious but critical issue under the MNAR assumption. We say a model is not identifiable when two different sets of parameters lead to the same model. Even in a simple case with a parametric missingness mechanism model (Wang et al., 2014; Miao et al., 2016), the model is generally not fully identifiable without imposing some extra assumptions. In the past decade, there has been tremendous development in proposing different model identifiability conditions in the literature. In this paper, we impose the following conditions to achieve the model identifiability under MNAR assumption.

(A1) Covariate $\mathbf{X}$ can be decomposed as $\mathbf{X} = (\mathbf{U}^{\mathrm{T}}, Z)^{\mathrm{T}}$ where Z is a scalar and $\mathbf{U}$ is of p-dimension, and the missingness mechanism model satisfies

$$\mathrm{pr}(R = 1 \mid y, \mathbf{x}) = \mathrm{pr}(R = 1 \mid y, \mathbf{u}) = \pi(y, \mathbf{u}). \tag{2.1}$$

The true missingness mechanism model is denoted as $\pi_0(y, \mathbf{u})$, and we assume that there exists a small positive constant $\epsilon > 0$ such that $\pi_0(y, \mathbf{u}) > \epsilon$ almost surely.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

(A2) Denote the support of Y,$\mathbf{U}$ and Z as $\mathcal{C} = \{(y, \mathbf{u}, z) : y \in \mathcal{Y}, \mathbf{u} \in \mathcal{U}, z \in \mathcal{Z}\}$. Assume $\mathcal{U}$ and $\mathcal{Z}$ are both bounded sets. Assume that given $R = 1$, random variable Y still preserves the support $\mathcal{Y}$.

(A3) Assume that there exists a bounded subset of $\mathcal{Z}$, denoted by $\mathcal{I}$, which contains non-empty interior points of $\mathcal{Z}$, such as, if $f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta}_1) = f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta}_2)$ on $\mathcal{C}_z = \{(y, \mathbf{u}, z) : y \in \mathcal{Y}, \mathbf{u} \in \mathcal{U}, z \in \mathcal{I}\}$, then $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$. Assume the corresponding Fisher information matrix restricted on $\mathcal{I}$, which is given by

$$E\left[\{\partial \log f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}\}\{\partial \log f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}\}^{\mathrm{T}} I(z \in \mathcal{I})\right],$$

is positive definite, where $I(\cdot)$ is the indicator function and $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$.

Based on conditions (A1)-(A3), our result below, with its proof in the supplementary material, shows that $\boldsymbol{\beta}$ and $\pi(y, \mathbf{u})$ are both identifiable.

Lemma 1. Under conditions (A1)-(A3), $\boldsymbol{\beta}$ and $\pi(y, \mathbf{u})$ are both identifiable.

We now provide some intuitions and interpretations of this result. In condition (A1), the variable Z is called the shadow variable in the literature. This condition implies that part of the covariate, denoted by Z, is independent of the missingness indicator R, conditional on the response Y

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

and the remaining covariates $\mathbf{U}$. The shadow variable assumption is popularly used in the literature, to name a few, Shao and Zhao (2013); Wang et al. (2014); Zhao and Shao (2015); Miao and Tchetgen Tchetgen (2016), etc., and is found to be useful in a variety of applications. In our context, we require Z to be univariate. If $\mathbf{Z}$ is multivariate, we recommend selecting one of its components as the shadow variable, as this allows for a more flexible modeling of the missingness mechanism (2.1).

Condition (A2) specifies the finite support of covariates $\mathbf{U}$ and Z. This is usually true empirically because many variables in social and biomedical studies, e.g., age, income, blood pressure, socioeconomic status, etc., have bounded ranges. In particular, $\mathbf{U}$ and Z can be any type of variables, e.g., continuous, categorical, ordinal, etc., as long as the finite support condition is satisfied. Also, condition (A2) excludes some extreme situations, e.g., subjects with observed responses only contain one type of subjects for the binary response case.

For condition (A3), one can verify that it satisfies if $f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})$ belongs to the exponential family with full rank and $\mathcal{I}$ is a closed interval. It is worthwhile to note that Condition (A3) is new. It is different from existing work on identifiability conditions which usually impose the completeness assumption; see, e.g., Zhao and Ma (2022); Miao et al. (2024).

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

In addition, we would like to emphasize that $\mathbf{U}$ could be an empty set but the shadow variable Z cannot. To illustrate, consider the simple linear regression where $f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})$ is the density of the normal distribution with mean $\beta_0 + \beta_1 z$ and variance 1 and the missingness mechanism model follows $\pi(y) = \text{expit}(\alpha_0 + \alpha_1 y)$ with $\text{expit}(\cdot) = \exp(\cdot)/\{1 + \exp(\cdot)\}$. Then, when the truth of the coefficient β_1 is nonzero, one can follow Lemma 1 to show that both $\boldsymbol{\beta}$ and $\pi(y)$ are identifiable. However, if the truth of β_1 equals zero; i.e., Z becomes an empty set, it is easy to create a counterexample as below. Consider two different sets of parameter values: $\beta_0 = 0$, $\alpha_0 = -0.5$, $\alpha_1 = 1$ and $\beta_0' = 1$, $\alpha_0' = 0.5$, $\alpha_1' = -1$. It is a simple calculation that both sets of parameter values result in the same likelihood function. Hence, it is not feasible to identify $\boldsymbol{\beta}$ and $\pi(y)$ in this case.

2.2 Likelihood Approximation

The conditional probability distribution function of (R, Y) given $\mathbf{X}$ from one single observation is

$$p(r, y, \mathbf{x}; \boldsymbol{\beta}, \pi) = \{\pi(y, \mathbf{u}) f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})\}^r \left\{1 - \int f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta}) \pi(t, \mathbf{u}) dt\right\}^{1-r}.$$

In Section 2.1, we showed that, under some conditions, the model is identifiable. That is, if $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$ and $\pi_1(y, \mathbf{u}) \neq \pi_2(y, \mathbf{u})$, then we must have two different densities $p(r, y, \mathbf{x}; \boldsymbol{\beta}_1, \pi_1) \neq p(r, y, \mathbf{x}; \boldsymbol{\beta}_2, \pi_2)$. We introduce

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

the log-likelihood function $\mathcal{L}(\boldsymbol{\beta}, \pi) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}, \pi)$, where $l_i(\boldsymbol{\beta}, \pi)$ equals $l(\boldsymbol{\beta}, \pi) \equiv \log\{p(r, y, \mathbf{x}; \boldsymbol{\beta}, \pi)\}$ applied to the ith subject $(r_i, y_i, \mathbf{x}_i)$, and

$$\begin{aligned} l(\boldsymbol{\beta}, \pi) =& r\log\pi(y, \mathbf{u}) + r\log f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta}) + \\ & (1-r)\log\left\{1 - \int f_{Y|\mathbf{X}}(t, \mathbf{x}; \boldsymbol{\beta})\pi(t, \mathbf{u})dt\right\}, \end{aligned}$$

is the corresponding log-likelihood from one single observation.

Our objective is to derive the efficient estimator of $\boldsymbol{\beta}$ via maximizing the log-likelihood function $\mathcal{L}(\boldsymbol{\beta}, \pi)$, which is semiparametric in the sense that the main interest is parameter $\boldsymbol{\beta}$ and the nuisance is the nonparametric function $\pi(y, \mathbf{u})$. We use the method of sieves (Grenander, 1981; Shen, 1997) to approximate the nonparametric nuisance $\pi(y, \mathbf{u})$.

Specifically, we use the B-spline basis (Schumaker, 2007) to construct the approximating functions. We first introduce a general B-spline set-up for a function of one-dimensional random variable X on $[0, 1]$. We introduce an extended partition on the interval $[0, 1]$, given by

$$\begin{aligned} \Delta \equiv& \{t_{-m+1} = \cdots = t_{-1} = 0 = t_0 < \\ & t_1 < \cdots < t_{K_n} < t_{K_n+1} = 1 = t_{K_n+2} = \cdots = t_{K_n+m}\} \end{aligned}$$

where m is the order the spline basis, and K_n is the number of interior knots. The interior knots $\{t_j, j = 1, \ldots, K_n\}$ can be chosen as evenly spaced partitions in $[0, 1]$ with length $1/(K_n + 1)$. Let $\{B_l^m(x)\}_{l=-m+1}^{K_n}$ be a one-dimensional normalized B-spline basis of order m associated with Δ. We

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

construct $B_l^m(x)$ from the recursive formula

$$B_l^m(x) = \frac{x - t_l}{t_{l+m-1} - t_l} B_l^{m-1}(x) + \frac{t_{l+m} - x}{t_{l+m} - t_{l+1}} B_{l+1}^{m-1}(x), l = -m+1, \ldots, K_n,$$

where $B_l^1(x) = I(t_l \le x \le t_{l+1})$, $l = 0, \ldots, K_n$.

After proper scaling and centering, we can assume that each component of $\mathbf{u}$ and y (based on observed data) has support on $[0, 1]$. Then we approximate the $(p+1)$-dimensional function logit $\pi(y, \mathbf{u})$ as

$$\begin{aligned} \text{logit } \pi(y, \mathbf{u}) = \eta(y, \mathbf{u}) &= \sum_{l_1, \ldots, l_{p+1} = -m+1}^{K_n} \tau_{l_1, \ldots, l_{p+1}} B_{l_1}^m(y) B_{l_2}^m(u_1) \cdots B_{l_{p+1}}^m(u_p) \\ &\equiv \mathbf{B}^m(y, \mathbf{u})^{\mathrm{T}} \boldsymbol{\tau}, \end{aligned} \tag{2.2}$$

where $\text{logit}(t) \equiv \log\{t/(1-t)\}$. Here $\mathbf{B}^m(y, \mathbf{u})$ and $\boldsymbol{\tau}$ are $s_n = (K_n + m)^{p+1}$-dimensional. They are constructed by stacking $\left\{ B_{l_1}^m(y) B_{l_2}^m(u_1) \cdots B_{l_{p+1}}^m(u_p) \right\}$ and $\left\{ \tau_{l_1, \ldots, l_{p+1}} \right\}$ as vectors respectively for all the combinations of the indices $l_1, \ldots, l_{p+1} \in \{-m+1, \ldots, K_n\}$. Also, the order of indices in $\boldsymbol{\tau}$ and $\mathbf{B}^m(y, \mathbf{u})$ should be the same:

$$\mathbf{B}^m(y, \mathbf{u}) = \begin{bmatrix} \vdots \\ B_{l_1}^m(y) B_{l_2}^m(u_1) \cdots B_{l_{p+1}}^m(u_p) \\ \vdots \end{bmatrix} \quad \boldsymbol{\tau} = \begin{bmatrix} \vdots \\ \tau_{l_1, l_2, \ldots, l_{p+1}} \\ \vdots \end{bmatrix}.$$

After the approximation of the propensity score function, our objective

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

becomes to maximize

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, \eta(y, \mathbf{u})) \;=\; & \sum_{i=1}^{n} r_i \log f_{Y|\mathbf{X}}(y_i, \mathbf{x}_i; \boldsymbol{\beta}) + \\ & \sum_{i=1}^{n} r_i \eta(y, \mathbf{u}) - \sum_{i=1}^{n} r_i \log\left\{1 + \exp\left(\eta(y, \mathbf{u})\right)\right\} + \\ & \sum_{i=1}^{n} (1 - r_i) \log\left\{1 - \int f_{Y|\mathbf{X}}(t, \mathbf{x}_i; \boldsymbol{\beta}) \mathrm{expit}\left(\eta(y, \mathbf{u})\right) dt\right\},\end{aligned}$$

or, equivalently,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\tau}) \;=\; & \sum_{i=1}^{n} r_i \log f_{Y|\mathbf{X}}(y_i, \mathbf{x}_i; \boldsymbol{\beta}) + \sum_{i=1}^{n} r_i \mathbf{B}^m(y_i, \mathbf{u}_i)^{\mathrm{T}} \boldsymbol{\tau} - \\ & \sum_{i=1}^{n} r_i \log\left\{1 + \exp\left(\mathbf{B}^m(y_i, \mathbf{u}_i)^{\mathrm{T}} \boldsymbol{\tau}\right)\right\} + \\ & \sum_{i=1}^{n} (1 - r_i) \log\left\{1 - \int f_{Y|\mathbf{X}}(t, \mathbf{x}_i; \boldsymbol{\beta}) \mathrm{expit}\left(\mathbf{B}^m(t, \mathbf{u}_i)^{\mathrm{T}} \boldsymbol{\tau}\right) dt\right\}\end{aligned} \tag{2.3}$$

over the sieve space

$$\mathcal{S}(m, K_n, M_n) = \left\{ (\boldsymbol{\beta}, \eta(y, \mathbf{u})) : \|\boldsymbol{\beta}\|_2 \le B_0, \sum_{l_1, \ldots, l_{p+1} = -m+1}^{K_n} |\tau_{l_1, \ldots, l_{p+1}}| \le M_n, \right\},$$

where $\|\boldsymbol{\beta}\| = \sqrt{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\beta}}$, B_0 is a given bound, K_n and M_n are some constants depending on the sample size n with their choices discussed later. With these bounds, we can guarantee that the sieve space $\mathcal{S}(m, K_n, M_n)$ is a compact set in a finite-dimensional space.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

3. Theory

Let $(\widehat{\boldsymbol{\beta}}_n, \widehat{\pi}_n)$ be the estimates achieving the maximum in (2.3). We now provide several asymptotic results, including consistency, asymptotic normality, and efficiency, for the estimates of the regression coefficients.

We first introduce some notations to ease our presentation. We denote the space $\mathcal{D} = [0,1]^{p+1}$. We denote $W^{k,\infty}(\mathcal{D})$ as a Sobolev space consisting of the functions defined on $\mathcal{D}$ with bounded kth order derivatives. Specifically when $k = 1$, the Sobolev norm of $f(y, \mathbf{u})$ is defined as $\|f\|_{W^{1,\infty}} = \|f\|_\infty + \|\nabla f\|_\infty$, where $\|f\|_\infty = \inf\{C \geq 0 : |f(y, \mathbf{u})| \leq C \text{ for almost every } (y, \mathbf{u})\}$, and ∇f is the gradient vector of f with respect to both y and $\mathbf{u}$.

We need the following conditions to establish our asymptotic theory, besides model identifiability conditions (A1)-(A3).

(A4) For a known integer $k > 1.5p + 4.5$, $\pi_0(y, \mathbf{u}) \in W^{k,\infty}(\mathbb{R}^{p+1})$.

(A5) $M_n = O(\log \log n)$, and $K_n = O(n^\alpha)$ with $(4k)^{-1} < \alpha < (6p+7)^{-1}$.

(A6) If there exists a constant vector $\mathbf{w}$ such that

$$[\partial \log\{f_{Y|\mathbf{X}}(y, \mathbf{u}, z_1; \boldsymbol{\beta}_0)/f_{Y|\mathbf{X}}(y, \mathbf{u}, z_2; \boldsymbol{\beta}_0)\}/\partial \boldsymbol{\beta}]^{\mathrm{T}} \mathbf{w} = 0,$$

for any $(y, \mathbf{u}, z_i) \in \mathcal{P} = \{(y, \mathbf{u}, z) : \pi(y, \mathbf{u}) > 0\}$, $i = 1, 2$, then $\mathbf{w} = \mathbf{0}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

Condition (A4) pertains to the smoothness of the unknown missingness mechanism model, and condition (A5) specifies the size of the sieve space $\mathcal{S}(m, K_n, M_n)$. Condition (A6) ensures that the score function for $\boldsymbol{\beta}$, denoted by $l_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \pi)$, is of full rank on $\mathcal{P}$. We present three theorems concerning the asymptotic properties of the parameter estimates $\boldsymbol{\beta}$ and the nuisance $\pi(y, \mathbf{u})$. The first theorem is on the consistency, with its proof in the supplementary material.

Theorem 1. Under conditions (A1)-(A6), $\widehat{\boldsymbol{\beta}}_n$ is a consistent estimator of the true parameter $\boldsymbol{\beta}_0$, and $\|\widehat{\pi}_n(y, \mathbf{u}) - \pi_0(y, \mathbf{u})\|_{W^{1,\infty}} \xrightarrow{p} 0$, when $n \to \infty$.

To obtain the asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$, we need to obtain a tighter bound for the convergence rate of the estimates, which is stated below, with its proof in the supplementary material.

Theorem 2. Under conditions (A1)-(A6), $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 \leq O_p(K_n^{-2k}) + o_p(n^{-1/2})$, and $\|\widehat{\pi}_n(y, \mathbf{u}) - \pi_0(y, \mathbf{u})\|_{L_2}^2 \leq O_p(K_n^{-2k}) + o_p(n^{-1/2})$.

Furthermore, we can now establish the asymptotic normality and semiparametric efficiency of $\widehat{\boldsymbol{\beta}}_n$, with its proof in the supplementary material.

Theorem 3. Under conditions (A1)-(A6), $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \equiv \mathbf{P}\left\{\boldsymbol{\phi}(\boldsymbol{\beta}_0, \pi_0)\boldsymbol{\phi}(\boldsymbol{\beta}_0, \pi_0)^{\mathrm{T}}\right\}$ is the semiparametric efficiency bound for $\boldsymbol{\beta}_0$ and $\boldsymbol{\phi}(\boldsymbol{\beta}_0, \pi_0)$ is the influence function of $\widehat{\boldsymbol{\beta}}_n$ defined in the Supplement.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

Here $\mathbf{P}$ is the expectation of the $\mathbf{P}_n$, which is the empirical measure based on the n i.i.d. observations. Thus, $\widehat{\boldsymbol{\beta}}_n$ is the semiparametrically efficient estimator of $\boldsymbol{\beta}_0$.

4. Numerical Implementation

This section presents a novel EM-type algorithm to estimate the parameter $\boldsymbol{\beta}$ via maximizing the objective function (2.3). Then, we propose a variance estimation method using the profile likelihood idea (Murphy and Van der Vaart 2000).

4.1 Parameter Estimation

Firstly, the complete-data log-likelihood (pretending that there were no missing values) is

$$\begin{aligned}&\sum_{i=1}^{n} R_i\{\log\pi(Y_i,\mathbf{U}_i)+\log f_{Y|\mathbf{X}}(Y_i,\mathbf{X}_i;\boldsymbol{\beta})\}+\\&\sum_{i=1}^{n}(1-R_i)\left\{\log\left[1-\pi(Y_i,\mathbf{U}_i)\right]+\log f_{Y|\mathbf{X}}(Y_i,\mathbf{X}_i;\boldsymbol{\beta})\right\}.\end{aligned}\tag{4.1}$$

We approximate the missing mechanism $\pi(Y,\mathbf{U})$ using B-spline basis functions as described in (2.2), then $\pi(Y_i,\mathbf{U}_i)$ in (4.1) can be replaced by $\boldsymbol{\tau}^T\mathbf{B}^m(Y_i,\mathbf{U}_i)$, where $\boldsymbol{\tau}$ and $\mathbf{B}^m(y,\mathbf{u})$ are vectors containing all permutations of $\tau_{l_1,\ldots,l_{p+1}}$ and $B^m_{l_1}(y)B^m_{l_2}(u_1)\cdots B^m_{l_{p+1}}(u_p)$ for $l_1,\ldots,l_{p+1}=1,\ldots,p+1$, respectively.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

After replacing $\pi(Y_i, \mathbf{U}_i)$, the unknown parameters in (4.1) are $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$. In the E-step, we need to compute the expectation of the second term of (4.1), that is

$$\sum_{i=1}^{n}(1-R_i)\left\{\log\left[1-\pi(Y_i,\mathbf{U}_i)\right]+\log f_{Y|\mathbf{X}}(Y_i,\mathbf{X}_i;\boldsymbol{\beta})\right\}, \tag{4.2}$$

given the parameter updates $\widehat{\boldsymbol{\beta}}^{(t)}$ and $\widehat{\boldsymbol{\tau}}^{(t)}$ for those subjects with missing values $R_i = 0$. With $\widehat{\boldsymbol{\beta}}^{(t)}$ and $\widehat{\boldsymbol{\tau}}^{(t)}$, the distribution of Y_i conditional on $R_i = 0$ and $\mathbf{X}_i$ is given by

$$\begin{aligned}\mathrm{pr}(Y \mid \mathbf{X}_i, R_i=0; \widehat{\boldsymbol{\beta}}^{(t)}, \widehat{\boldsymbol{\tau}}^{(t)}) &= \frac{\{1-\pi(Y,\mathbf{U}_i)\}f_{Y|\mathbf{X}}(Y,\mathbf{X}_i;\widehat{\boldsymbol{\beta}}^{(t)})}{\int\{1-\pi(s,\mathbf{U}_i)\}f_{Y|\mathbf{X}}(s,\mathbf{X}_i;\widehat{\boldsymbol{\beta}}^{(t)})ds} \\ =&\frac{[1+\exp\{(\widehat{\boldsymbol{\tau}}^{(t)})^{\mathrm{T}}\mathbf{B}^m(Y,\mathbf{U}_i)\}]^{-1}f_{Y|\mathbf{X}}(Y,\mathbf{X}_i;\widehat{\boldsymbol{\beta}}^{(t)})}{\int[1+\exp\{(\widehat{\boldsymbol{\tau}}^{(t)})^{\mathrm{T}}\mathbf{B}^m(s,\mathbf{U}_i)\}]^{-1}f_{Y|\mathbf{X}}(s,\mathbf{X}_i;\widehat{\boldsymbol{\beta}}^{(t)})ds}.\end{aligned} \tag{4.3}$$

By taking expectation of (4.2) with respect to the distribution in (4.3), we obtain the objective function $Q(\boldsymbol{\beta}, \boldsymbol{\tau})$ for the M-step

$$\begin{aligned}Q(\boldsymbol{\beta},\boldsymbol{\tau}) =& \sum_{i=1}^{n} R_i \log f_{Y|\mathbf{X}}(Y_i,\mathbf{X}_i;\boldsymbol{\beta}) + \sum_{i=1}^{n}(1-R_i)\widehat{E}_i\{\log f_{Y|\mathbf{X}}(Y,\mathbf{X}_i;\boldsymbol{\beta})\}+ \\ & \sum_{i=1}^{n} R_i[\boldsymbol{\tau}^{\mathrm{T}}\mathbf{B}^m(Y_i,\mathbf{U}_i) - \log\{1+\exp(\boldsymbol{\tau}^{\mathrm{T}}\mathbf{B}^m(Y_i,\mathbf{U}_i))\}]- \\ & \sum_{i=1}^{n}(1-R_i)\widehat{E}_i[\log\{1+\exp(\boldsymbol{\tau}^{\mathrm{T}}\mathbf{B}^m(Y,\mathbf{U}_i))\}],\end{aligned} \tag{4.4}$$

where

$$\widehat{E}_i\{\log f_{Y|\mathbf{X}}(Y,\mathbf{X}_i;\boldsymbol{\beta})\} =$$

$$\frac{\int \log f_{Y|\mathbf{X}}(y,\mathbf{X}_i;\boldsymbol{\beta})[1+\exp\{(\widehat{\boldsymbol{\tau}}^{(t)})^{\mathrm{T}}\mathbf{B}^m(y,\mathbf{U}_i)\}]^{-1}f_{Y|\mathbf{X}}(y,\mathbf{X}_i;\widehat{\boldsymbol{\beta}}^{(t)})dy}{\int[1+\exp\{(\widehat{\boldsymbol{\tau}}^{(t)})^{\mathrm{T}}\mathbf{B}^m(s,\mathbf{U}_i)\}]^{-1}f_{Y|\mathbf{X}}(s,\mathbf{X}_i;\widehat{\boldsymbol{\beta}}^{(t)})ds}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

and

$$\widehat{E}_i[\log\{1+\exp(\boldsymbol{\tau}^{\mathrm{T}}\mathbf{B}^m(Y,\mathbf{U}_i))\}] = \frac{\int \log\{1+\exp(\boldsymbol{\tau}^{\mathrm{T}}\mathbf{B}^m(y,\mathbf{U}_i))\}[1+\exp\{(\widehat{\boldsymbol{\tau}}^{(t)})^{\mathrm{T}}\mathbf{B}^m(y,\mathbf{U}_i)\}]^{-1} f_{Y|\mathbf{X}}(y,\mathbf{X}_i;\widehat{\boldsymbol{\beta}}^{(t)})dy}{\int [1+\exp\{(\widehat{\boldsymbol{\tau}}^{(t)})^{\mathrm{T}}\mathbf{B}^m(s,\mathbf{U}_i)\}]^{-1} f_{Y|\mathbf{X}}(s,\mathbf{X}_i;\widehat{\boldsymbol{\beta}}^{(t)})ds}.$$

For the first two terms of the objective function in (4.4), the only unknown parameters are $\boldsymbol{\beta}$; for the last two terms in (4.4), the only unknown parameters are $\boldsymbol{\tau}$. Then we can obtain the update $\widehat{\boldsymbol{\beta}}^{(t+1)}$ by maximizing the sum of the first two terms, which is denoted by

$$Q(\boldsymbol{\beta}) \equiv \sum_{i=1}^{n} R_i \log f_{Y|\mathbf{X}}(Y_i,\mathbf{X}_i;\boldsymbol{\beta}) + \sum_{i=1}^{n}(1-R_i)\widehat{E}_i\{\log f_{Y|\mathbf{X}}(Y,\mathbf{X}_i;\boldsymbol{\beta})\}; \quad (4.5)$$

and we can obtain the update $\boldsymbol{\tau}^{(t+1)}$ by maximizing the sum of the last two terms, which is denoted by

$$\begin{aligned} Q(\boldsymbol{\tau}) \equiv \quad & \sum_{i=1}^{n} R_i[\boldsymbol{\tau}^{\mathrm{T}}\mathbf{B}^m(Y_i,\mathbf{U}_i) - \log\{1+\exp(\boldsymbol{\tau}^{\mathrm{T}}\mathbf{B}^m(Y_i,\mathbf{U}_i))\}] \\ & - \sum_{i=1}^{n}(1-R_i)\widehat{E}_i[\log\{1+\exp(\boldsymbol{\tau}^{\mathrm{T}}\mathbf{B}^m(Y,\mathbf{U}_i))\}]. \end{aligned} \quad (4.6)$$

We start with initial values $\widehat{\boldsymbol{\beta}}^{(0)} = \widehat{\boldsymbol{\beta}}_{\mathrm{MAR}}$ and $\widehat{\boldsymbol{\tau}}^{(0)} = \mathbf{0}$ and repeat the EM iterations until convergence ($\widehat{\boldsymbol{\beta}}_{\mathrm{MAR}}$ is the naive estimator of $\boldsymbol{\beta}$ only using data from the completely observed subjects).

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

4.2 M-Step

We can maximize (4.5) and (4.6) by first computing the derivatives and then use the Newton-Raphson algorithm. But when $Y \sim f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})$ has a normal distribution, we can simplify this problem using the Gauss-Hermite quadrature and solve the optimization using existing, well-studied, robust, algorithms.

For the objective function (4.5), the expectation $\widehat{E}_i\{\log f_{Y|\mathbf{X}}(Y, \mathbf{X}_i; \boldsymbol{\beta})\}$ can be written as

$$\widehat{E}_i\{\log[f_{Y|\mathbf{X}}(Y, \mathbf{X}_i; \boldsymbol{\beta})]\} = C_i \int \log[f_{Y|\mathbf{X}}(y, \mathbf{X}_i; \boldsymbol{\beta})] h_i(y) f_{Y|\mathbf{X}}(y, \mathbf{X}_i; \widehat{\boldsymbol{\beta}}^{(t)}) dy, \tag{4.7}$$

where

$$1/C_i \equiv \int [1 + \exp\{(\widehat{\boldsymbol{\tau}}^{(t)})^{\mathrm{T}} \mathbf{B}^m(s, \mathbf{U}_i)\}]^{-1} f_{Y|\mathbf{X}}(s, \mathbf{X}_i; \widehat{\boldsymbol{\beta}}^{(t)}) ds,$$

and

$$h_i(y) \equiv [1 + \exp\{(\widehat{\boldsymbol{\tau}}^{(t)})^{\mathrm{T}} \mathbf{B}^m(y, \mathbf{U}_i)\}]^{-1}.$$

Let υ_i for $i = 1, \ldots, G$ be the roots of the Hermite polynomial of degree G

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

and w_i be the corresponding weights, then (4.7) can be approximated with

$$\begin{aligned}
&\widehat{E}_i\{\log f_{Y|\mathbf{X}}(Y, \mathbf{X}_i; \boldsymbol{\beta})\} \\
&\approx \frac{C_i}{\sqrt{\pi}} \sum_{j=1}^{G} w_j h_i(\sqrt{2}\widehat{\sigma}^{(t)} \upsilon_j + \widehat{\mu}^{(t)}) \log f_{Y|\mathbf{X}}(\sqrt{2}\widehat{\sigma}^{(t)} \upsilon_j + \widehat{\mu}^{(t)}, \mathbf{X}_i; \boldsymbol{\beta}) \\
&\equiv \sum_{j=1}^{G} w_{ij}^* \log f_{Y|\mathbf{X}}(\sqrt{2}\widehat{\sigma}^{(t)} \upsilon_j + \widehat{\mu}^{(t)}, \mathbf{X}_i; \boldsymbol{\beta})
\end{aligned} \tag{4.8}$$

where $(\widehat{\mu}^{(t)}, \widehat{\sigma}^{(t)})$ are the location and scale parameters, which can be computed from $\widehat{\boldsymbol{\beta}}^{(t)}$ and

$$w_{ij}^* \equiv \frac{1}{\sqrt{\pi}} C_i w_j h_i(\sqrt{2}\widehat{\sigma}^{(t)} \upsilon_j + \widehat{\mu}^{(t)}).$$

The approximation in (4.8) leads to

$$\begin{aligned}
Q(\boldsymbol{\beta}) \approx &\sum_{i=1}^{n} R_i \log f_{Y|\mathbf{X}}(Y_i, \mathbf{X}_i; \boldsymbol{\beta}) + \\
&\sum_{i=1}^{n} \sum_{j=1}^{G} (1 - R_i) w_{ij}^* \log f_{Y|\mathbf{X}}(\sqrt{2}\widehat{\sigma}^{(t)} \upsilon_j + \widehat{\mu}^{(t)}, \mathbf{X}_i; \boldsymbol{\beta}),
\end{aligned} \tag{4.9}$$

which is the weighted log-likelihood function of a sample with no missing values. Thus, maximizing $Q(\boldsymbol{\beta})$ becomes finding the MLE in a weighted linear regression problem, which can be done using existing software (e.g., R function glm()).

Similarly, for the expectation $\widehat{E}_i[\log\{1 + \exp(\boldsymbol{\tau}^{\mathrm{T}} \mathbf{B}^m(Y, \mathbf{U}_i))\}]$ in (4.6), we can rewrite it as

$$\begin{aligned}
&\widehat{E}_i[\log\{1 + \exp(\boldsymbol{\tau}^{\mathrm{T}} \mathbf{B}^m(Y, \mathbf{U}_i))\}] \\
=&C_i \int \log\left\{1 + \exp[\boldsymbol{\tau}^T \mathbf{B}^m(y, \mathbf{U}_i)]\right\} h_i(y) f_{Y|\mathbf{X}}(y, \mathbf{X}_i; \widehat{\boldsymbol{\beta}}^{(t)}) dy,
\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

thus we can further approximate the objective function $Q(\boldsymbol{\tau})$ with

$$Q(\boldsymbol{\tau}) \approx \sum_{i=1}^{n} R_i[\boldsymbol{\tau}^{\mathrm{T}}\mathbf{B}^m(Y_i, \mathbf{U}_i) - \log\{1 + \exp(\boldsymbol{\tau}^{\mathrm{T}}\mathbf{B}^m(Y_i, \mathbf{U}_i))\}] - \sum_{i=1}^{n}\sum_{j=1}^{G}(1 - R_i)w_{ij}^{*}\log\left\{1 + \exp[\boldsymbol{\tau}^{T}\mathbf{B}^m(\sqrt{2}\widehat{\sigma}^{(t)}\upsilon_j + \widehat{\mu}^{(t)}, \mathbf{U}_i)]\right\}. \tag{4.10}$$

The approximated objective function (4.10) is the weighted log-likelihood function of logistic regression with complete samples; thus, updating $\boldsymbol{\tau}$ is equivalent to computing the MLE for a weighted logistic regression problem, which can also be done using existing software. When $Y \sim f_{Y|\mathbf{X}}(y, \mathbf{x}; \boldsymbol{\beta})$ does not have a normal distribution, it is still possible to use the importance sampling technique and a normal distribution as the proposal distribution. However, we will not discuss this topic in this paper. Lastly, we would discuss choosing the order and the number of knots using the B-spline basis. More splines capture complex patterns and reduce bias but can lead to overfitting or an unstable estimation procedure. Techniques like AIC and BIC can help find the sweet spot between bias and variance. Performing a sensitivity analysis, where we evaluate model performance with different choices, can further refine the selection and assess the impact.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

4.3 Estimating the Limit of the Covariance Matrix

Recall that $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\tau})$ in (2.3) is log-likelihood function given the data. The profile log-likelihood function for $\boldsymbol{\beta}$ is $pl(\boldsymbol{\beta}) = \max_{\boldsymbol{\tau}} \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\tau})$. In general, the profile likelihood theory in Murphy and Van der Vaart (2000) holds for the profile likelihood $pl(\boldsymbol{\beta})$. Thus, we can estimate the limit of the covariance matrix of $\widehat{\boldsymbol{\beta}}_n$ using the negative inverse of the Hessian matrix of $pl(\widehat{\boldsymbol{\beta}}_n)$.

We can compute the Hessian matrix of $pl(\widehat{\boldsymbol{\beta}}_n)$ numerically, and the (k, l)th element is computed by

$$\epsilon_n^{-2} \left\{ pl(\widehat{\boldsymbol{\beta}}_n + \mathbf{e}_k\epsilon_n + \mathbf{e}_l\epsilon_n) - pl(\widehat{\boldsymbol{\beta}}_n + \mathbf{e}_k\epsilon_n) - pl(\widehat{\boldsymbol{\beta}}_n + \mathbf{e}_l\epsilon_n) + pl(\widehat{\boldsymbol{\beta}}_n) \right\}, \tag{4.11}$$

where $\mathbf{e}_k$ is the kth canonical vector, and ϵ_n is a constant of the order $n^{-1/2}$. We have found that setting ϵ_n between $1/\sqrt{n}$ and $5/\sqrt{n}$ can usually yield satisfactory results. Since the purpose of ϵ_n is used to approximate certain derivatives numerically, there are also some automatic differentiation methods from computing literature that could be used (e.g., Ridders 1982).

Except for $pl(\widehat{\boldsymbol{\beta}}_n) = \mathcal{L}(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\tau}})$, the other three terms inside the bracket of (4.11) (i.e., $pl(\widehat{\boldsymbol{\beta}}_n + \mathbf{e}_k\epsilon_n + \mathbf{e}_l\epsilon_n)$, $pl(\widehat{\boldsymbol{\beta}}_n + \mathbf{e}_k\epsilon_n)$, and $pl(\widehat{\boldsymbol{\beta}}_n + \mathbf{e}_l\epsilon_n)$) need additional computation. The computation of the profile likelihood is similar to using the EM-based algorithm described in Section 4.1. The only

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

difference is that, in the EM algorithm, $\boldsymbol{\beta}$ is fixed at its initial value $\widehat{\boldsymbol{\beta}}^{(0)}$ and we only update $\boldsymbol{\tau}$ until convergence.

5. Simulation Studies

We investigate the performance of the proposed method in this section. We consider three different simulation settings and compare the proposed method with the pseudo-likelihood method (Tang et al. 2003). The pseudo-likelihood method estimate the regression parameters by maximizing the pseudo-likelihood

$$\prod_{i=1}^{n} \frac{R_i f_{Y|\mathbf{X}}(Y_i|\mathbf{X}_i;\boldsymbol{\beta})}{\int f_{Y|\mathbf{X}}(Y_i|\mathbf{x};\boldsymbol{\beta}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}},$$

where $f_{\mathbf{x}}(\mathbf{x})$ is the marginal density of $\mathbf{x}$. The marginal distribution of $\mathbf{X}$, which is usually unknown, can be estimated or replaced with its empirical distribution instead. We also compare with the estimator based on the MAR assumption (i.e., only using the fully observed data) and the oracle estimator (i.e., using all simulated data) for comparisons. The sample size is set to be 1000 and the Monte Carlo replication is 2000 in all three simulations.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

5.1 Simulation Settings

In the first simulation setting, the one-dimensional covariate X follows a standard normal distribution $\text{Norm}(0, 1)$, and we generate the response Y using the following conditional distribution $Y \mid X \sim \text{Norm}(1 + X, 1)$. The missing data mechanism model only depends on Y and is given as

$$\text{logit}\{\pi(Y)\} = 1 - Y + 0.5Y^2.$$

In this setting, the true parameter value is $\boldsymbol{\theta}_1 = (\beta_0, \beta_1, \sigma) = (1, 1, 1)$ (i.e., parameters from $\text{Norm}(1 + X, 1)$) and around 25% of the responses Y are missing. When applying the proposed method, we use the linear bases (i.e., $q = 2$), and the number of interior knots is $b_n = 2$.

For the second simulation setting, there are two one-dimensional covariates $\mathbf{X}^{\mathrm{T}} = [Z, U]$: $Z \sim \text{Norm}(0, 1)$ and $U \mid Z \sim \text{Norm}(1 - Z, 1)$, where Z is the shadow variable. The regression model generating Y is $Y \mid \mathbf{X} \sim \text{Norm}(-1 + 4Z - 2U, 1)$. The missing data mechanism depends on both Y and U and is given as

$$\text{logit}\{\pi(Y, U)\} = 4 - 3Y + U.$$

The true parameter value is $\boldsymbol{\theta}_2 = (\beta_0, \beta_1, \beta_2, \sigma) = (-1, 4, -2, 1)$ and around 25% of the responses Y are missing. We use the linear bases $q = 2$ with

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

three interior knots $b_n = 3$ when applying the proposed method.

The third simulation setting has three one-dimensional covariates $\mathbf{X} = (Z, U_1, U_2)$, where $Z \sim \text{Norm}(0, 1)$ and $U_1, U_2 \mid Z \sim \text{Norm}(1 - Z, 1)$. The regression model generating the response Y is $Y \mid \mathbf{X} \sim \text{Norm}(-1+4Z-U_1-U_2, 1)$. Again, Z is the shadow variable, and the missing data mechanism only depends on Y and $\boldsymbol{U} = (U_1, U_2)$ as

$$\text{logit}\{\pi(Y, \boldsymbol{U})\} = 4 - 3Y + U_1 + U_2.$$

The true parameter value is $\boldsymbol{\theta}_3 = (\beta_0, \beta_1, \beta_2, \beta_3, \sigma) = (-1, 4, -1, -1, 1)$ and the proportion of missing data is around 23%. We use the linear bases $q = 2$ with four interior knots $b_n = 4$ when using the proposed method.

5.2 Simulation Results

The results of estimating the main parameter of interest in these three simulation settings are summarized in Tables 1–3. Of the four methods, "Oracle" refers to the estimator using all simulated data, "MAR" denotes the estimator only using the observations without missing data, "Pseudo" is the pseudo-likelihood method, and "proposed" corresponds to the method we propose in this paper. We compare the bias, empirical standard deviation (SD), mean of standard errors (SE) (of all the replicates), and the

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

coverage probability (CP).

The Oracle estimator has the smallest standard error and standard deviation in all three simulation settings, as it uses more information than other estimators. The MAR estimator, not surprisingly, fails in all three simulation settings. The pseudo-likelihood method is consistently the most conservative regarding the coverage probability. Our proposed method has a smaller standard error and standard deviation than the pseudo-likelihood method, confirming our theoretical estimation efficiency results. The pseudo-likelihood method generally has a smaller bias than the proposed method. However, it is worth noting that the comparisons in settings 2 and 3 are not fair. This is because the pseudo-likelihood method needs the conditional distribution of Z given $\mathbf{U}$ (or (U_1, U_2)), and we use the correct parametric models in both settings when applying the pseudo-likelihood method. The proposed method, however, does not require such model specifications.

In addition to parameter estimation, we also investigate the performance of the method of sieves. In Figure 1, which is for simulation setting 1, the solid line is the true missing data mechanism $\pi(Y)$, the dash-dotted line is the median of the estimated $\pi(Y)$ from 2000 Monte Carlo samples,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

and the dashed lines are the 97.5% and 2.5% quantiles of the 2000 estimated $\pi(Y)$. Our proposed method has good performance between $Y = 1$ and $Y = 2$, but when Y is small, the method of sieves tends to overestimate the value of $\pi(Y)$. To improve the estimation of $\pi(Y)$, we can increase the number of interior knots and use bases of higher order (e.g., $q = 3$); consequently, a larger sample size n is needed.

Our proposed method demonstrates good properties even with a limited sample size. The proposed method has a smaller standard error and requires fewer model assumptions than the pseudo-likelihood method.

6. Real Data Application

We apply the proposed method to the Medical Information Mart for Intensive Care III database (MIMIC-III, Johnson et al. 2016). MIMIC-III is an open EHR dataset containing information on anonymous intensive care unit patients admitted to Beth Israel Deaconess Medical Center from 2001 to 2012, and this dataset has been widely used in academic and industrial research. This dataset includes but is not limited to information such as vital signs, medication, laboratory tests, and demographics.

When we pre-processed the dataset, we found that while most biomark-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

Parameter	Method	Bias	SD	SE	CP
β_0	Oracle	-0.0010	0.0321	0.0316	0.9515
	MAR	-0.0012	0.0386	0.0379	0.9500
	Pseudo	-0.0012	0.0410	0.0513	0.9850
	Proposed	-0.0006	0.0401	0.0399	0.9490
β_1	Oracle	-0.0007	0.0320	0.0316	0.9490
	MAR	0.0884	0.0350	0.0362	0.3055
	Pseudo	0.0005	0.0423	0.0612	0.9930
	Proposed	0.0118	0.0373	0.0376	0.9405
σ	Oracle	-0.0016	0.0223	0.0223	0.9495
	MAR	0.0416	0.0266	0.0267	0.6690
	Pseudo	-0.0016	0.0249	0.0333	0.9895
	Proposed	0.0017	0.0249	0.0246	0.9470

Table 1: Simulation Setting 1: results for estimating the $\boldsymbol{\beta}$ and σ.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

Parameter	Method	Bias	SD	SE	CP
β_0	Oracle	-0.0004	0.0456	0.0448	0.9490
	MAR	-0.1530	0.0609	0.0619	0.3000
	Pseudo	-0.0010	0.0643	0.0669	0.9625
	Proposed	0.0172	0.0608	0.0605	0.9410
β_1	Oracle	-0.0003	0.0458	0.0448	0.9445
	MAR	-0.0727	0.0554	0.0547	0.7345
	Pseudo	0.0025	0.0580	0.0613	0.9590
	Proposed	-0.0093	0.0545	0.0540	0.9440
β_2	Oracle	0.0005	0.0321	0.0317	0.9390
	MAR	0.0418	0.0377	0.0379	0.8050
	Pseudo	0.0014	0.0388	0.0415	0.9640
	Proposed	-0.0112	0.0370	0.0373	0.9380
σ	Oracle	-0.0021	0.0227	0.0223	0.9405
	MAR	-0.0117	0.0257	0.0255	0.9180
	Pseudo	-0.0036	0.0271	0.0266	0.9350
	Proposed	-0.0011	0.0262	0.0259	0.9400

Table 2: Simulation Setting 2: results for estimating $\boldsymbol{\beta}$ and σ.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

Parameter	Method	Bias	SD	SE	CP
β_0	Oracle	0.0019	0.0556	0.0549	0.9470
	MAR	-0.1456	0.0719	0.0717	0.4675
	Pseudo	-0.0002	0.0769	0.0808	0.9590
	Proposed	0.0349	0.0726	0.0704	0.9175
β_1	Oracle	-0.0028	0.0553	0.0549	0.9530
	MAR	-0.0683	0.0651	0.0651	0.8155
	Pseudo	0.0009	0.0689	0.0764	0.9680
	Proposed	-0.0162	0.0646	0.0643	0.9405
β_2	Oracle	-0.0021	0.0314	0.0317	0.9520
	MAR	0.0199	0.0364	0.0366	0.9205
	Pseudo	-0.0011	0.0380	0.0411	0.9675
	Proposed	-0.0134	0.0360	0.0361	0.9360
β_3	Oracle	0.0003	0.0322	0.0317	0.9470
	MAR	0.0226	0.0372	0.0366	0.9050
	Pseudo	0.0016	0.0387	0.0411	0.9610
	Proposed	-0.0106	0.0367	0.0361	0.9345
σ	Oracle	-0.0024	0.0220	0.0223	0.9480
	MAR	-0.0113	0.0247	0.0253	0.9280
	Pseudo	-0.0026	0.0256	0.0265	0.9580
	Proposed	-0.0014	0.0251	0.0257	0.9560

Table 3: Simulation Setting 3: results for estimating the main parameter

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

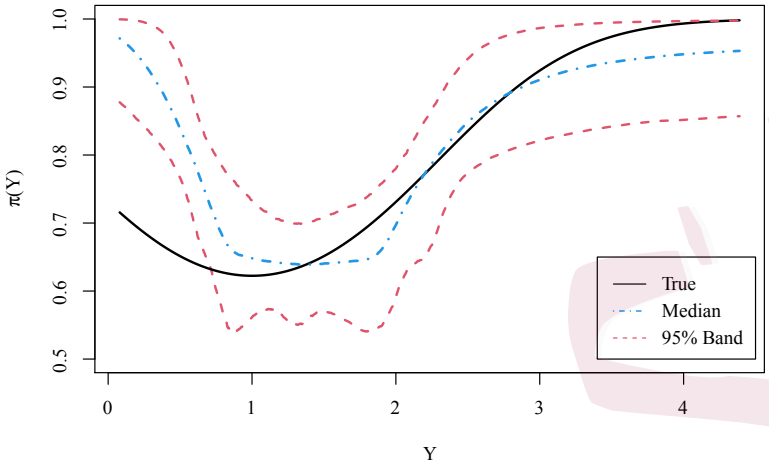

Figure 1: Simulation Setting 1: results for estimating the nonparametric missingness mechanism model $\pi(y)$. The dashed lines are 95% ad hoc confidence bands.

ers have a missing rate of less than 3%, the albumin level in the blood sample has a missing rate of around 30%. In the EHR database, data collection depends on the patients' visiting process (i.e., the measurements are only available when the patients need such examination). The patient's health status can be a potential confounding factor for the patient's visiting process and the biomarker measurement. Thus, whether a missing biomarker could depend on its own value and the more general MNAR assumption is reasonable.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

In this real data application, we aim to study the association between the albumin level and other variables under the MNAR assumption via a linear regression model. Except for the albumin level, other variables have no missing values. Biomedical studies showed that the calcium level is highly correlated with the albumin level (e.g., Katz and Klotz 1953, Butler et al. 1984). Therefore, we use the calcium level as the shadow variable. In addition, we follow Zhao and Chen (2020) and select three other biomarkers as explanatory variables: red blood cell, magnesium, and the sequential organ failure assessment score (SOFA). Our data has sample size $n = 1359$, where 421 of the observations have missing albumin levels.

The analysis results are summarized in Table 4. Though the proposed method does not alter the significance of any of the variables, the parameter estimates from the proposed method are very different from the method based on the MAR assumption. The pseudo-likelihood method yields similar results as the proposed method. But the standard errors are larger than those of the proposed method except for the coefficient of magnesium. The proposed method suggests that the naive analysis based on subjects with fully observed data underestimates the albumin level's association with red blood cells and calcium (positive direction) and SOFA (negative direction).

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

More importantly, the proposed method indicates that the association between the albumin level and magnesium should be negative instead of positive.

Metric	(Intercept)	red blood cell	magnesium	SOFA	calcium
MAR					
Estimate	-1.328	0.125	0.021	-0.061	0.171
SE	0.084	0.009	0.016	0.005	0.010
95% CI	[-1.494, -1.163]	[0.107, 0.144]	[-0.053, 0.011]	[-0.070, -0.052]	[0.152, 0.190]
Proposed					
Estimate	-8.455	0.611	-0.135	-0.349	0.804
SE	0.189	0.037	0.117	0.023	0.013
95% CI	[-8.824, -8.085]	[0.538, 0.685]	[-0.363, 0.094]	[-0.394, -0.303]	[0.778, 0.831]
Pseudo					
Estimate	-9.177	0.629	-0.145	-0.317	0.890
SE	0.597	0.056	0.085	0.027	0.061
95% CI	[-10.347, -8.006]	[0.518, 0.739]	[-0.312, 0.023]	[-0.372, -0.265]	[0.769, 1.010]

Table 4: Real Data Application: Comparing the parameter estimate (Estimate), its standard error (SE), and 95% confidence interval (CI) using the completely observed data (MAR), pseudo-likelihood method (Pseudo), and using the proposed method (Proposed).

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

Supplementary Material

In the online supplementary material, Section S1 provides proofs for all the lemmas and theorems in the main paper, and Section S2 provides more details on the Gauss-Hermite Quadrature used in the main paper.

Acknowledgments

Tian is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2023-03479. Zeng is supported in part by U.S. National Institutes of Health (R01HL173128). Zhao is supported in part by U.S. National Science Foundation (DMS 1953526, 2122074 and 2310942), U.S. National Institutes of Health (R01DC021431) and the American Family Funding Initiative of UW-Madison.

References

Butler, S., R. Payne, I. Gunn, J. Burns, and C. Paterson (1984). Correlation between serum ionised calcium and serum albumin concentrations in two hospital populations. *Br Med J (Clin Res Ed) 289*(6450), 948–950.

Chen, X., H. Hong, and A. Tarozzi (2008). Semiparametric efficiency in gmm models with auxiliary data. *The Annals of Statistics 36*, 808–843.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

REFERENCES

Grenander, U. (1981). *Abstract Inference*. New York: Wiley.

Johnson, A. E., T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark (2016). Mimic-iii, a freely accessible critical care database. *Scientific data 3*(1), 1–9.

Katz, S. and I. M. Klotz (1953). Interactions of calcium with serum albumin. *Archives of biochemistry and biophysics 44*(2), 351–361.

Kim, J. K. and C. L. Yu (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association 106*(493), 157–165.

Li, M., Y. Ma, and J. Zhao (2022). Efficient estimation in a partially specified nonignorable propensity score model. *Computational Statistics & Data Analysis 174*, 107322.

Little, R. J. A. and D. B. Rubin (2002). *Statistical analysis with missing data* (2 ed.). Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York.

Miao, W., P. Ding, and Z. Geng (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association 111*(516), 1673–1683.

Miao, W., L. Liu, Y. Li, E. J. Tchetgen Tchetgen, and Z. Geng (2024). Identification and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *ACM/JMS Journal of Data Science 1*(2), 1–23.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

REFERENCES

Miao, W. and E. J. Tchetgen Tchetgen (2016). On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika 103*(2), 475–482.

Murphy, S. A. and A. W. Van der Vaart (2000). On profile likelihood. *Journal of the American Statistical Association 95*(450), 449–465.

Ridders, C. (1982). Accurate computation of F′(x) and F′(x)F′′(x). *Advances in Engineering Software 4*(2), 75–76.

Schumaker, L. (2007). *Spline Functions: Basic Theory*. Cambridge University Press.

Shao, J. and J. Zhao (2013). Estimation in longitudinal studies with nonignorable dropout. *Statistics and Its Interface 6*, 303–313.

Shen, X. (1997). On methods of sieves and penalization. *The Annals of Statistics 25*(6), 2555–2591.

Shetty, S., Y. Ma, and J. Zhao (2023). The pursuit of efficiency versus robustness: A learning experience from analyzing a semiparametric nonignorable propensity score model. *Observational Studies 9*(1), 97–104.

Shetty, S., Y. Ma, and J. Zhao (2025). Robust estimation under a semiparametric propensity model for nonignorable missing data. *Electronic Journal of Statistics 19*(1), 956–981.

Tang, G., R. J. Little, and T. E. Raghunathan (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika 90*(4), 747–764.

Tang, N., P. Zhao, and H. Zhu (2014). Empirical likelihood for estimating equations with

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

nonignorably missing data. *Statistica Sinica 24*, 723–747.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer Series in Statistics. Springer, New York.

Wang, D. and S. X. Chen (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics 37*, 490–517.

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

Zhao, J. and C. Chen (2020). A nuisance-free inference procedure accounting for the unknown missingness with application to electronic health records. *Entropy 22*(10), 1154.

Zhao, J. and Y. Ma (2018). Optimal pseudolikelihood estimation in the analysis of multivariate missing data with nonignorable nonresponse. *Biometrika 105*(2), 479–486.

Zhao, J. and Y. Ma (2022). A versatile estimation procedure without estimating the nonignorable missingness mechanism. *Journal of the American Statistical Association 117*(540), 1916–1930.

Zhao, J. and J. Shao (2015). Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association 110*(512), 1577–1590.

Zhao, J. and J. Shao (2017). Approximate conditional likelihood for generalized linear models with general missing data mechanism. *Journal of Systems Science and Complexity 30*(1),

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0204

139–153.

Statistica Sinica