Statistical Papers (2024) 65:3235–3259
https://doi.org/10.1007/s00362-023-01522-0

REGULAR ARTICLE

Semiparametric estimation in generalized additive partial linear models with nonignorable nonresponse data

Jierui Du[1] · Xia Cui[1]

Received: 4 June 2023 / Revised: 1 December 2023 / Published online: 30 December 2023

Abstract

We address the semiparametric challenge of identifying and estimating generalized additive partial linear models with nonignorable missingness in the response. Identifiability is ensured under instrumental variable assumption that there is an instrumental covariate related to the prospensity but unrelated to the response variable, or the assumption that the conditional score function is linear in the response variable. We propose a new estimating equation for the prospensity by taking expectation of the unobservable part on a linear combination of all covariates rather than the covariates themselves. This estimating equation does not suffer from the typical curse of dimensionality. Then the unknown nonparametric function is approximated by polynomial spline basis functions and we construct estimating equations for mean of response based on the inverse probability weighting. Under some regular conditions, we establish asymptotic normality of the proposed estimators for parametric components and consistency of the estimators of nonparametric functions. Simulation studies demonstrate that the proposed inference procedure performs well in many settings. The proposed method is applied to analyze the household income dataset from the Chinese Household Income Project Survey 2013.

Keywords Generalized additive partial linear models · Nonignorable missingness · Identifiability · Instrumental variable · Asymptotic normality

Mathematics Subject Classification 62D10 · 62G08

✉ Xia Cui
cuixia@gzhu.edu.cn

Jierui Du
dujierui@gzhu.edu.cn

[1] School of Economics and Statistics, Guangzhou University, Guangzhou 510006, Guangdong, People's Republic of China

1 Introduction

Missing data is a prevalent issue in research areas like biomedicine, social sciences, and survey sampling. Underlying any missing data problem is the statistical model for the data if none of the data were missing (Tsiatis 2006). The missingness mechanism plays a crucial role in distinguishing different types of missingness problems. The missingness is named ignorable if it depends on the observed data only; otherwise, it is named nonignorable (Little and Rubin 2019; Zhao and Ma 2022). In practice, generalized additive partial linear models, combining interpretability and flexibility, are widely used for modeling different response types. Nonignorable models are underused due to the complexity of the identification and estimation procedures needed to recover parameters of interest as functions of observed data (Nabi and Bhattacharya 2022).

Identification is generally not accessible under nonignorable missingness without additional assumptions. One approach is to employ the shadow variable strategy (Wang et al. 2014; Zhao and Shao 2015; Miao and Tchetgen Tchetgen 2016). Another similar method involves using instrumental variables (Tchetgen Tchetgen and Wirth 2017; Sun et al. 2018). Recently, Zhao and Ma (2022) and Li et al. (2022) combined these two approaches, establishing clear identifiability for the model. However, selecting suitable instrumental variables or shadow variables can be difficult, especially when dealing with numerous covariates (Cameron and Trivedi 2005). An alternative approach to address identifiability without using instrumental variables or shadow variables relies on stronger assumptions regarding the distribution of the response or the response mechanism. Stronger assumptions about the response mechanism allow for the derivation of identifiability based on the distribution of the observed data, as demonstrated by Morikawa and Kim (2021) and Beppu et al. (2022). Miao et al. (2016), Cui et al. (2017), and Du et al. (2023) made assumptions that the response in the full data follows a specific distribution, such as exponential families. However, when these stronger assumptions on the response distribution and mechanism may lead to misspecification, utilizing instrumental variables or shadow variables remains a reasonable approach.

Further advancements are needed to develop estimation methods when the observed likelihood is identifiable. Extensive research has been conducted in this area, with various approaches proposed. For example, Wang et al. (2014), Shao and Wang (2016), and Wang et al. (2021) employed the generalized estimating equations approach. The empirical likelihood approach was utilized by Tang et al. (2014) and Cui et al. (2022). Calibration was employed by Kott and Chang (2010) and Hamori et al. (2019), while the pseudo likelihood approach was applied by Fang and Shao (2016) and Chen et al. (2021). These studies contribute to the existing literature by providing different estimation methods for addressing this issue.

Limited attention has been given in the existing literature to situations where regression models involve nonparametric functions of interest and the response is affected by nonignorable missingness, despite its prevalence in practical applied research. Du et al. (2023) tackle the challenge of identifying and estimating generalized additive partial linear models by assuming that the response in the full data follows exponential family. On the other hand, Shao and Wang (2022) propose estimators for regression models with a single nonparametric function when the data distribution is unknown, but their focus does not specifically address model identifiability. In this paper, generalized additive partial linear models are identified through the imposition of three types of monotone missing data mechanisms: logistic model, probit model, and complementary log-log model. The logistic and probit models are popular missing data mechanisms (Wang et al. 2014). The complementary log-log model has an important application in the area of survival analysis and hazard modeling (An and Brown 2008). These three models are likely to be most familiar to the target audience. Polynomial spline basis functions are used to approximate the unknown nonparametric function, and estimating equations for the mean response are formulated based on inverse probability weighting. Our contributions focus on three main aspects.

(1) Our proposed approach identifies generalized additive partial linear models through the imposition of three types of monotone missing data mechanisms: logistic model, probit model, and complementary log-log model. Identifiability is achieved by either assuming instrumental variable dependence without additional assumptions or assuming a linear relationship between the score function and the response variable, without the use of instrumental variables. The mild sufficient conditions for identifiability stem from leveraging the analytical properties of the propensity function.
(2) The missing model parameter estimators are obtained using the conditional score function. To address the curse of dimensionality, we employ dimension reduction techniques to achieve easily attainable univariate kernel estimation. The parameter and nonparametric function estimators in the regression model are obtained using inverse probability weighting. The unknown smooth functions are approximated by a linear combination of regression splines and incorporated into the covariate vector for statistical inference using generalized estimation equations.
(3) Under certain regularity conditions, we establish the asymptotic normality of the proposed estimators for the parametric components and the convergence rate of the estimators for the nonparametric functions. Simulation studies demonstrate the favorable performance of the proposed inference procedure across various settings. We also apply the proposed method to a dataset from the Chinese Household Income Project study conducted in 2013.

The paper is structured as follows. Section 2 establishes the sufficient conditions for the identifiability of the observed likelihood in the generalized additive partially linear models under nonignorable missingness. In Sect. 3, we introduce the estimation procedure and establish the consistency and asymptotic normality of the estimators. The performance of the proposed method is evaluated through simulation studies in Sect. 4. Section 5 demonstrates the application of the new method using data from the Chinese Household Income Project 2013. Concluding remarks are provided in Sect. 6. The proofs of Theorem 1-4 can be found in the Supplementary Material.

2 Model settings and identifiability

Let Y be the response variable, $\mathbf{X}$ and $\mathbf{Z}$ be the fully observed covariates, where $Y \in \mathbb{R}$, $\mathbf{X} \equiv (1, X_1, \ldots, X_{d_1-1})^\top \in \mathbb{R}^{d_1}$ and $\mathbf{Z} \equiv (Z_1, \ldots, Z_{d_2})^\top \in \mathbb{R}^{d_2}$. Define the binary variable r to be the missingness indicator, if Y is observable, r takes 1, otherwise takes 0. We assume that the probability $P(r = 1|Y = y, \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$ depends on y, $\mathbf{x}$, and $\mathbf{z}$ and denote it by $\pi(y, \mathbf{x}, \mathbf{z}; \alpha, \boldsymbol{\theta})$. We specify it using a logistic model, probit model or complementary log-log model as follows

$$\pi(y, \mathbf{x}, \mathbf{z}; \alpha, \boldsymbol{\theta}) = \text{expit}(\xi) \text{ or } \Phi(\xi) \text{ or } 1 - \exp\{-\exp(\xi)\}, \tag{1}$$

where $\xi = \alpha y + \boldsymbol{\theta}^\top(\mathbf{x}^\top, \mathbf{z}^\top)^\top$, $\text{expit}(\cdot) \equiv \exp(\cdot)/\{1 + \exp(\cdot)\}$, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. $\alpha \in \mathbb{R}$ is the nonignorable parameter and $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_{d_1+d_2-1})^\top$ is an unknown $(d_1 + d_2)$-dimensional parameters. In the model described by Eq. (1), the probability of missingness depends on the potentially missing Y through the parameter α. When $\alpha = 0$, the missing mechanism is independent of the potential missing Y, indicating that it is missing at random. Conversely, if $\alpha \neq 0$, it indicates nonignorable missingness.

Denoting $p(y|\mathbf{x}, \mathbf{z})$ as the conditional density function of y given $\mathbf{x}$ and $\mathbf{z}$, the conditional density function of a single sample based on observed data can be expressed as

$$\begin{aligned}&\{p(y, r = 1|\mathbf{x}, \mathbf{z})\}^{I(r=1)}\{p(r = 0|\mathbf{x}, \mathbf{z})\}^{I(r=0)}\\ &= \{p(r = 1|y, \mathbf{x}, \mathbf{z})p(y|\mathbf{x}, \mathbf{z})\}^{I(r=1)}[E\{p(r = 0|Y, \mathbf{X}, \mathbf{Z})|\mathbf{X}, \mathbf{Z}\}]^{I(r=0)}.\end{aligned}$$

Suppose we have an independent random sample $(Y_i, r_i, \mathbf{X}_i, \mathbf{Z}_i)$, $i = 1, \cdots, n$. The observed likelihood given $\mathbf{X}_i, \mathbf{Z}_i$ can be written as

$$\begin{aligned}L_n &= \prod_{i=1}^{n} \Big\{\pi(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \alpha, \boldsymbol{\theta}) p(Y_i|\mathbf{X}_i, \mathbf{Z}_i)\Big\}^{r_i}\\ &\quad\times\Big[\int \{1 - \pi(y, \mathbf{X}_i, \mathbf{Z}_i; \alpha, \boldsymbol{\theta})\} p(y|\mathbf{X}_i, \mathbf{Z}_i) dy\Big]^{1-r_i}.\end{aligned} \tag{2}$$

Nonignorable missingness in Y poses challenges to the identifiability of the observed likelihood, as highlighted by Wang et al. (2014). In Sect. 4, we demonstrate an unidentifiable example and discuss the resulting fluctuations in the estimators when the model lacks identifiability. Identifiability of the observed likelihood function (2) depends on the unique determination of $\pi(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \alpha, \boldsymbol{\theta})$ and $p(Y_i|\mathbf{X}_i, \mathbf{Z}_i)$ given $\mathbf{X}_i, \mathbf{Z}_i$. If there exist two sets of parameters $(\alpha, \boldsymbol{\theta}, p(y|\mathbf{x}, \mathbf{z}))$ and $(\alpha^*, \boldsymbol{\theta}^*, p^*(y|\mathbf{x}, \mathbf{z}))$ such that

$$\pi(y, \mathbf{x}, \mathbf{z}; \alpha, \boldsymbol{\theta}) p(y|\mathbf{x}, \mathbf{z}) = \pi(y, \mathbf{x}, \mathbf{z}; \alpha^*, \boldsymbol{\theta}^*) p^*(y|\mathbf{x}, \mathbf{z}),$$

holds for all $(y, \mathbf{x}, \mathbf{z})$ in an open set of $\mathbb{R}^{d_1+d_2+1}$, taking logarithms on both sides gives

$$h(\xi) + \log p(y|\mathbf{x}, \mathbf{z}) = h(\xi^*) + \log p^*(y|\mathbf{x}, \mathbf{z}), \tag{3}$$

where $\xi^* = \alpha^* y + \boldsymbol{\theta}^{*\top}(\mathbf{x}^\top, \mathbf{z}^\top)^\top$, $h(\xi)$ can take three types of forms $\log\{\text{expit}(\xi)\}$, $\log\{\Phi(\xi)\}$, $\log[1 - \exp\{-\exp(\xi)\}]$. The observed likelihood is identifiable if (3) implies that

$$\alpha = \alpha^*, \quad \boldsymbol{\theta} = \boldsymbol{\theta}^*, \quad p(y|\mathbf{x}, \mathbf{z}) = p^*(y|\mathbf{x}, \mathbf{z}).$$

To ensure identifiability, we can adopt the instrumental variable assumption, as defined in Assumption 1, following a similar approach as proposed by Tchetgen Tchetgen and Wirth (2017) and Sun et al. (2018).

Assumption 1 The missing mechanism $\pi(Y, \mathbf{X}, \mathbf{Z}; \alpha, \boldsymbol{\theta})$ includes a variable that is conditionally independent of Y given the other covariates.

The absence of a direct effect of the instrumental variable on the response echoes an assumption commonly encountered in causal inference. Moreover, in the context of this paper, the instrumental variable is integrated within the missing mechanism model. Tchetgen Tchetgen and Wirth (2017) introduced homogeneous additive selection bias, making selection bias independent of the instrumental variable, which enables the identifiability of mean functionals for observed covariates. Sun et al. (2018) restricted the ratio $p(y|\mathbf{x}, \mathbf{z})/p^*(y|\mathbf{x}, \mathbf{z})$ and establishes the identifiability of $p(y, r|\mathbf{x}, \mathbf{z})$. By employing three distinct forms of monotone missing data mechanisms: logistic model, probit model, and complementary log-log model, we are able to leverage their analytical attributes to streamline the requirements for ensuring model identifiability.

Theorem 1 *Under Assumption 1, if there is at least one continuous variable in the nonlinear component, the observed likelihood (2) is identifiable.*

The proof is provided in the Supplementary Material. Theorem 1 establishes a sufficient condition for the identifiability of the observed likelihood under the instrumental variable assumption. However, determining a reasonable instrumental variable beforehand is often impractical, and detecting its presence from observed data is challenging. In cases where the instrumental variable assumption fails or reasonable instrumental variables are difficult to choose, stronger assumptions on the response may be necessary beyond what standard statistical methods typically require.

Assumption 2 Let $\mu(\mathbf{X}, \mathbf{Z}) = E(Y|\mathbf{X}, \mathbf{Z})$, and $\upsilon(\cdot)$ represents the nuisance parameter. Defining $S_\mu\{\mathbf{X}, \mathbf{Z}; \mu, \upsilon(\cdot)\} = \partial \log p(Y|\mathbf{X}, \mathbf{Z}; \mu, \upsilon(\cdot))/\partial\mu$, and we allow:

$$S_\mu\{\mathbf{X}, \mathbf{Z}; \mu, \upsilon(\cdot)\} = a\{\mathbf{X}, \mathbf{Z}; \mu, \upsilon(\cdot)\}Y + b\{\mathbf{X}, \mathbf{Z}; \mu, \upsilon(\cdot)\}. \tag{4}$$

We now illustrate three cases demonstrating the validity of Assumption 2 for various common distributions.

Example 1 (Exponential family case): Assuming that the probability density function $p(Y|\mathbf{X}, \mathbf{Z})$ belongs to the exponential family, then

$$S_\mu(\mathbf{X}, \mathbf{Z}) = \frac{Y - \mu}{E[\{Y - \mu)\}^2|\mathbf{X}, \mathbf{Z}]},$$

and $a\{\mathbf{X}, \mathbf{Z}; \mu, \upsilon(\cdot)\} = 1/E\{(Y-\mu)^2|\mathbf{X}, \mathbf{Z}\}$ and $b\{\mathbf{X}, \mathbf{Z}; \mu, \upsilon(\cdot)\} = -\mu(\mathbf{X}, \mathbf{Z})/E\{(Y-\mu)^2|\mathbf{X}, \mathbf{Z}\}$. The exponential family encompasses various common distributions, such as the normal distribution and gamma distribution for continuous responses, the Bernoulli distribution for binary responses, and the Poisson distribution and geometric distribution for discrete responses.

Example 2 (Quasi-likelihood case): For the quasi-Poisson model with nonignorable nonresponse data, we can specify the structure of the probability density function based on assumptions about the conditional mean and variance functions in the following manner:

$$p(Y|\mathbf{X}, \mathbf{Z}; \mu, \upsilon(\cdot)) = \frac{\exp\{(Y \log \mu - \mu)/\phi)\}}{E[\exp\{(Y \log \mu - \mu)/\phi)\}|\mathbf{X}, \mathbf{Z}]},$$

then

$$S_\mu(\mathbf{X}, \mathbf{Z}) = \frac{Y - \mu}{\phi\mu},$$

and $a\{\mathbf{X}, \mathbf{Z}; \mu, \upsilon(\cdot)\} = 1/\{\phi\mu(\mathbf{X}, \mathbf{Z})\}$ and $b\{\mathbf{X}, \mathbf{Z}; \mu, \upsilon(\cdot)\} = -1/\phi$. This approach can be generalized to other quasi-likelihood statistical models as well.

Example 3 (Truncated distribution case): We assume the probability density function $p(Y|\mathbf{X}, \mathbf{Z}; \mu, \nu(\cdot))$ takes the following form:

$$p(Y|\mathbf{X}, \mathbf{Z}; \mu, \nu(\cdot)) = \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-(Y-\mu)^2}{2\sigma^2})}{\Phi(b - \mu/\sigma) - \Phi(a - \mu/\sigma)}, \qquad a \leq y \leq b,$$

where Φ is the cumulative distribution function of the standard normal distribution. Often the goal is to make inference back to the original population and not on the truncated population that is sampled (Hattaway 2010). In this case, the inference is focused on estimating μ, which represents the expectation of the original distribution. Let

$$S_\mu(\mathbf{X}, \mathbf{Z}) = \frac{Y - \mu}{\sigma^2} - \frac{\partial \log[\Phi\{(b - \mu)/\sigma) - \Phi((a - \mu)/\sigma\}]}{\partial\mu}, \qquad a \leq y \leq b,$$

where $a\{\mathbf{X}, \mathbf{Z}; \mu, \upsilon(\cdot)\} = 1/\sigma^2$ and $b\{\mathbf{X}, \mathbf{Z}; \mu, \upsilon(\cdot)\} = -\mu(\mathbf{X}, \mathbf{Z})/\sigma^2 - \partial \log[\Phi\{(b-\mu(\mathbf{X}, \mathbf{Z}))/\sigma) - \Phi((a - \mu(\mathbf{X}, \mathbf{Z}))/\sigma\}]/\partial\mu$. We can extend this approach to truncated distributions.

Let $\mu(\mathbf{X}, \mathbf{Z}) = \lambda(\eta)$, where $\lambda(\cdot)$ represents the inverse of the link function between the response and regression parameter η that is modeled as an additive partial linear function

$$\eta = \boldsymbol{\beta}^\top \mathbf{X} + \sum_{k=1}^{d_2} g_k(Z_k). \tag{5}$$

To ensure identifiability, we assume that the additive nonparametric functions in (5) are centered, i.e., $E[g_k(Z_k)] = 0$ for $k = 1, \ldots, d_2$. The inclusion of a linear component

$\boldsymbol{\beta}^\top \mathbf{X}$ in model (5) makes it easier to interpret, while the inclusion of the nonparametric component $\sum_{k=1}^{d_2} g_k(Z_k)$ enhances its flexibility.

Theorem 2 *Under Assumption 2, if the inverse of the link function $\lambda(\cdot)$ is a known one-to-one, first differential function and there is at least one continuous variable in the nonlinear component, then*

(i) *When Y is a binary variable, $\log \lambda(x)$ is strictly concave and the sign of the first derivative of the nonlinear component $\sum_{k=1}^{d_2} g_k(z_k)$ is known at point zero, the observed likelihood (2) is identifiable;*
(ii) *When Y is a discrete variable with at least three values, the observed likelihood (2) is identifiable if the sign of α is known;*
(iii) *When Y is a continuous variable and $h(\xi) = \log expit(\xi)$ is used, the observed likelihood (2) is identifiable when the sign of at least one element of the parameter vector $(\alpha, \boldsymbol{\theta}^\top)^\top$ is known;*
(iv) *When Y is a continuous variable, $h(\xi) = \log\{\Phi(\xi)\}$ or $\log[1-\exp\{-\exp(\xi)\}]$, the observed likelihood (2) is identifiable.*

The proof is provided in the Supplementary Material. The three examples above illustrate the wide applicability of Theorem 2. In contrast to Theorem 1 in Du et al. (2023), Theorem 2 expands the identifiability of the models range to a more general form and also facilitates the establishment of identifiable pseudo-likelihood functions, all without requiring instrumental variable assumptions. The inverse of the link function $\lambda(\cdot)$ is a known one-to-one, first differential function commonly employed in quasi-likelihood models (Wang et al. 2011). For binary variables, commonly used propensity functions, like the logistic model, probit model, and complementary log-log model, satisfy the condition that $\log \lambda(x)$ is strictly concave. Estimating $g_k'(x_{k0})$ involves using a local least squares algorithm, with x_{k0} chosen as a fixed point within a neighborhood where missingness does not occur (of length $O(n^{-1/5})$) (Fan et al. 1996). Prior knowledge of the sign of the unknown parameters in the missing mechanism models is required for parameter identifiability in the case of discrete response variables with at least three values and continuous response variables with a logistic missing data mechanism. According to Krosnick et al. (2002), factors like respondents' cognitive level, motivation, and social status influence nonresponse probability. Based on this, we can speculate on the trend of nonresponse probability and infer the sign of the parameters in the missing mechanism model. For instance, in a household income survey, high-income individuals might be less likely to disclose their true income, suggesting $\alpha < 0$. The identifiability of the observed likelihood (2) is guaranteed when the model (4) reduces to generalized additive models, as stated in Theorem 1 and Theorem 2.

3 Estimation method

By considering a nonparametric form of $p(y|\mathbf{x}, \mathbf{z})$ and utilizing the observed likelihood (2), The score function method proposed by Cui and Zhou (2017) for the parameters in the missing model can be derived as follows:

$$\sum_{i=1}^{n}\left\{r_i\frac{\pi'(Y_i,\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})}{\pi(Y_i,\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})}-(1-r_i)\frac{E[\{\pi'(Y_i,\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})\}|\mathbf{X}_i,\mathbf{Z}_i]}{E[\{1-\pi(Y_i,\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta}))\}|\mathbf{X}_i,\mathbf{Z}_i]}\right\}, \tag{6}$$

where $\boldsymbol{\delta}=(\alpha,\boldsymbol{\theta}^\top)^\top$, $\pi'(\cdot)$ denotes the partial derivative of $\pi(\cdot)$ with respect to $\boldsymbol{\delta}$. To estimate the parameters in equation (6) using the kernel method, a multivariate kernel is required. However, the standard nonparametric kernel regression estimators face challenges due to the curse of dimensionality. In this paper, we propose an improved formulation of equation (6) as:

$$\mathbf{V}(\boldsymbol{\delta})=\sum_{i=1}^{n}\left\{r_i\frac{\pi'(Y_i,\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})}{\pi(Y_i,\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})}+(1-r_i)\frac{\omega'(O_i;\boldsymbol{\delta})}{\omega(O_i;\boldsymbol{\delta})}\right\}, \tag{7}$$

where $\omega(O;\boldsymbol{\delta})=E[\{1-\pi(Y,\mathbf{X},\mathbf{Z};\boldsymbol{\delta})\}|O]$, $O=\boldsymbol{\theta}^\top(\mathbf{X}^\top,\mathbf{Z}^\top)^\top$, $\omega'(\cdot)$ denotes the partial derivative of $\omega(\cdot)$ with respect to $\boldsymbol{\delta}$. The score function (7) is unbiased since

$$\begin{aligned}E\{\mathbf{V}(\boldsymbol{\delta})\}&=\sum_{i=1}^{n}E\left\{r_i\frac{\pi'(Y_i,\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})}{\pi(Y_i,\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})}+(1-r_i)\frac{\omega'(O_i;\boldsymbol{\delta})}{\omega(O_i;\boldsymbol{\delta})}\right\}\\&=\sum_{i=1}^{n}E[E\{\pi'(Y_i,\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})+\omega'(\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})|O_i\}]\\&=0.\end{aligned}$$

Given observational data, we have

$$\begin{aligned}\omega(O_i;\boldsymbol{\delta})&=E[r_i\{1-\pi(Y_i,\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})\}/\pi(Y_i,\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})|O_i],\\\omega'(O_i;\boldsymbol{\delta})&=-E[r_i\pi'(Y_i,\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})/\pi(Y_i,\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})|O_i].\end{aligned}$$

There are some existing methods of estimating $\omega(O;\boldsymbol{\delta})$ and $\omega'(O;\boldsymbol{\delta})$, for details see Fan et al. (1996). The local constant estimators are given

$$\begin{aligned}\hat{\omega}(o;\boldsymbol{\delta})&=\frac{\sum_{i=1}^{n}k_h(O_i-o)[r_i\{1-\pi(Y_i,\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})\}/\pi(Y_i,\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})]}{\sum_{i=1}^{n}k_h(O_i-o)},\\\hat{\omega}'(o;\boldsymbol{\delta})&=-\frac{\sum_{i=1}^{n}k_h(O_i-o)\{r_i\pi'(Y_i,\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})/\pi(Y_i,\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})\}}{\sum_{i=1}^{n}k_h(O_i-o)},\end{aligned}$$

where $k(\cdot)$ is a given kernel function, h represents the bandwidth, and $k_h(t)$ is defined as $k(t/h)/h$.

Hence, we formulate the estimation equation for the parameter $\boldsymbol{\delta}$ in the missing model as follows:

$$\hat{\mathbf{V}}(\boldsymbol{\delta})=\sum_{i=1}^{n}\hat{\mathbf{V}}_i(\boldsymbol{\delta})=\sum_{i=1}^{n}\left\{r_i\frac{\pi'(Y_i,\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})}{\pi(Y_i,\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})}+(1-r_i)\frac{\hat{\omega}'(\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})}{\hat{\omega}(\mathbf{X}_i,\mathbf{Z}_i;\boldsymbol{\delta})}\right\}. \tag{8}$$

In estimating equation (8), we simplify the problem by projecting the propensity function onto a linear combination of covariates, transforming it into a univariate kernel estimation problem. While this approach may yield less efficient estimators, it offers simplicity and ease of implementation. In practical calculations, the nleqslv package in R can be used to solve equation (8) and obtain the estimator for $\boldsymbol{\delta}$.

The parameter $\boldsymbol{\delta}$ in the missing model exhibits consistency and asymptotic normality, as described in Theorem 3. This is contingent upon the asymptotic properties of the estimated function $\hat{\mathbf{V}}(\boldsymbol{\delta})$.

Theorem 3 *Under identifiable observed likelihood (2) and the satisfaction of conditions (A)-(E) in the Supplementary Material, if $nh^2/\log(1/h) \to \infty$ and $nh^4 \to 0$, the following holds:*

(i) $\hat{\boldsymbol{\delta}}$ *converges in probability to the true value* $\boldsymbol{\delta}_0$*,*

(ii) $\sqrt{n}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{\Omega}(\boldsymbol{\delta}_0))$*,*

where $\boldsymbol{\Omega}(\boldsymbol{\delta}) = A^{-1}(\boldsymbol{\delta})B(\boldsymbol{\delta})\{A^{-1}(\boldsymbol{\delta})\}^\top$,

$$A(\boldsymbol{\delta}) = E\{\partial \mathbf{V}_i(\boldsymbol{\delta})/\partial \boldsymbol{\delta}^\top\}, \quad B(\boldsymbol{\delta}) = E\{\mathbf{R}_i(\boldsymbol{\delta})\mathbf{R}_i^\top(\boldsymbol{\delta})\}, \tag{9}$$

and

$$\mathbf{R}_i(\boldsymbol{\delta}) = \frac{\{1 - r_i/\pi(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\delta})\}\omega'(\mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\delta})}{\omega(\mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\delta})}.$$

The proof of Theorem 3 is provided in the Supplementary Material. Based on Theorem 3, the asymptotic representation for $\hat{\mathbf{V}}(\boldsymbol{\delta})$ can be viewed as a special case of generalized estimating equations as described in Wang et al. (2021). In practical applications, the covariance matrix can be obtained using $\hat{A}^{-1}(\hat{\boldsymbol{\delta}})\hat{B}(\hat{\boldsymbol{\delta}})\{\hat{A}^{-1}(\hat{\boldsymbol{\delta}})\}^\top$, where

$$\hat{A}(\hat{\boldsymbol{\delta}}) = n^{-1}\sum_{i=1}^{n} \partial \hat{\mathbf{V}}_i(\hat{\boldsymbol{\delta}})/\partial \boldsymbol{\delta}^\top, \quad \hat{B}(\hat{\boldsymbol{\delta}}) = n^{-1}\sum_{i=1}^{n} \mathbf{R}_i(\hat{\boldsymbol{\delta}})\mathbf{R}_i^\top(\hat{\boldsymbol{\delta}}).$$

Now we will consider the estimators of unknown parameter vector $\boldsymbol{\beta}$ and unknown functions $g_k(Z_k) = 0, k = 1, \ldots, d_2$. When the propensity score is known, one approach to estimation is to plug the estimated $\hat{\boldsymbol{\delta}}$ from (8) into $\pi(y, \mathbf{x}, \mathbf{z}; \boldsymbol{\delta})$ and use inverse weighting of the complete cases. This method employs a generalized estimating equation and estimates the nonparametric components using polynomial splines. Recall that $\mathbf{Z} = (Z_1, \ldots, Z_{d_2})^\top$ represents a vector of covariates, $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{id_2})^\top$ is the vector of covariates for the ith observation, and η_i for the ith observation can be expressed as $\eta_i = \boldsymbol{\beta}^\top \mathbf{X}_i + \sum_{k=1}^{d_2} g_k(Z_{ik})$. Assuming that Z_k is distributed on a compact interval $[t_k^l, t_k^r], k = 1, \ldots, d_2$, without loss of generality, we can set all intervals to be $[t_k^l, t_k^r] = [0, 1], k = 1, \ldots, d_2$. According to the approach proposed by Wang and Yang (2007), the smooth unknown functions g_k's can be effectively approximated using a linear combination of polynomial spline functions. Let $\mathcal{S}_n$ be the space of polynomial splines on the interval [0, 1] of order $q \geq 1$. We define a knot sequence with J interior knots and denote it as

$$\tau_{-q} = \cdots = \tau_{-1} = \tau_0 = 0 < \tau_1 < \ldots < \tau_J < 1 = \tau_{J+1} = \cdots = \tau_{J+q+1},$$

where $J \equiv J_n$ is chosen to increase as the sample size n increases, and the specific order is provided in condition (I) in the Supplementary Material. Then $\mathcal{S}_n$ consists of functions $\tilde{\omega}$ that satisfy the following properties: (i) $\tilde{\omega}$ is a polynomial of degree q on each of the subintervals $I_s = [\tau_s, \tau_{s+1}), s = 0, \ldots, J_n - 1, I_{J_n} = [\tau_{J_n}, 1]$; (ii) for $q \geq 1$, $\tilde{\omega}$ is a $(q-1)$ times continuously differentiable on [0, 1]. For the kth covariate Z_k, let $\{\tilde{b}_{j,k}(Z_k), j = 1, \ldots, J_n + q + 2, k = 1, \ldots, d_2\}$ be the B-spline basis functions of order q of the space of $\mathcal{S}_n$. Let $N_n = J_n + q + 1$, we adopt the normalized B-spline space $\mathcal{S}_n^0$ introduced in Xue and Yang (2006) with the normalized basis as follows, $1 \leq j \leq N_n, 1 \leq k \leq d_2$,

$$B_{j,k}(Z_k) = \sqrt{N_n}\left\{\tilde{b}_{j+1,k}(Z_k) - \frac{E(\tilde{b}_{j+1,k})}{E(\tilde{b}_{1,k})}\tilde{b}_{1,k}(Z_k)\right\}. \tag{10}$$

The normalized B-spline approximation for $g_k(Z_k)$ can then be expressed as following

$$g_k(Z_k) \approx \tilde{g}_k(Z_k) = \sum_{j=1}^{N_n} \gamma_{j,k} B_{j,k}(Z_k) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N_n} \gamma_{j,k} B_{j,k}(Z_{ik}). \tag{11}$$

Denoting $\boldsymbol{\gamma} = (\gamma_{1,1}, \ldots, \gamma_{N_n,d_2})^\top$ as a vector of coefficients of dimension $N_n d_2$, $\mathbf{B}_{i,k} = (B_{1,k}(Z_{ik}), \ldots, B_{J_n,k}(Z_{ik}))^\top$ and $\mathbf{B}_i = (\mathbf{B}_{i,1}^\top, \ldots, \mathbf{B}_{i,d_2}^\top)^\top$, we can simplify the notation by representing $\sum_{j=1}^{N_n} \gamma_{j,k} B_{j,k}(Z_k)$ as $\tilde{g}_k(Z_k)$. Using the normalized B-splines, we approximate η_i as $\tilde{\eta}_i = \mathbf{X}_i^\top \boldsymbol{\beta} + \tilde{g}(\mathbf{Z}_i) = \mathbf{X}_i^\top \boldsymbol{\beta} + \mathbf{B}_i^\top \boldsymbol{\gamma}$.

Suppose that the conditional variance function $\text{var}(Y|\mathbf{X}, \mathbf{Z}) = \phi V(\mu(\mathbf{X}, \mathbf{Z}))$ for $\phi > 0$ and some known positive function V. Let the quasi-score function be defined as

$$q(\boldsymbol{\beta}, g) = \frac{Y - \mu}{V(\mu)} \times \frac{\partial \mu}{\partial \eta}, \tag{12}$$

By replacing the unknown smooth function with the approximation given in Equation (11), we can obtain the following estimating equation:

$$\sum_{i=1}^{n} \mathbf{U}_i(\hat{\boldsymbol{\delta}}, \boldsymbol{\beta}, \tilde{g}) = \begin{pmatrix} \sum_{i=1}^{n} \mathbf{U}_{\boldsymbol{\beta},i}(\hat{\boldsymbol{\delta}}, \boldsymbol{\beta}, \tilde{g}) \\ \sum_{i=1}^{n} \mathbf{U}_{\boldsymbol{\gamma},i}(\hat{\boldsymbol{\delta}}, \boldsymbol{\beta}, \tilde{g}) \end{pmatrix} = 0, \tag{13}$$

where

$$\mathbf{U}_{\boldsymbol{\beta},i}(\hat{\boldsymbol{\delta}}, \boldsymbol{\beta}, \tilde{g}) = \frac{r_i q_i(\boldsymbol{\beta}, \tilde{g})}{\pi_i(\hat{\boldsymbol{\delta}})}\mathbf{X}_i, \quad \mathbf{U}_{\boldsymbol{\gamma},i}(\hat{\boldsymbol{\delta}}, \boldsymbol{\beta}, \tilde{g}) = \frac{r_i q_i(\boldsymbol{\beta}, \tilde{g})}{\pi_i(\hat{\boldsymbol{\delta}})}\mathbf{B}_i.$$

Then the estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be obtained by solving for (13), and $\hat{g} = \mathbf{B}^\top \hat{\boldsymbol{\gamma}}$.

We will introduce some notation. Let

$$D(\boldsymbol{\delta}, \boldsymbol{\beta}, g) = E\{\partial \mathbf{U}_{\boldsymbol{\beta},i}(\boldsymbol{\delta}, \boldsymbol{\beta}, g)/\partial \boldsymbol{\beta}^\top\} - G(\boldsymbol{\delta}, \boldsymbol{\beta}, g) E\{\partial \mathbf{U}_{\boldsymbol{\gamma},i}(\boldsymbol{\delta}, \boldsymbol{\beta}, g)/\partial \boldsymbol{\beta}^\top\},$$

where

$$G(\boldsymbol{\delta}, \boldsymbol{\beta}, g) = E\{\partial \mathbf{U}_{\boldsymbol{\beta},i}(\boldsymbol{\delta}, \boldsymbol{\beta}, g)/\partial \boldsymbol{\gamma}^\top\}[E\{\partial \mathbf{U}_{\boldsymbol{\gamma},i}(\boldsymbol{\delta}, \boldsymbol{\beta}, g)/\partial \boldsymbol{\gamma}^\top\}]^{-1}.$$

And $\Sigma(\boldsymbol{\delta}, \boldsymbol{\beta}, g) = E\{m(\boldsymbol{\delta}, \boldsymbol{\beta}, g) m(\boldsymbol{\delta}, \boldsymbol{\beta}, g)^\top\}$, where

$$\begin{aligned} m_i(\boldsymbol{\delta}, \boldsymbol{\beta}, g) = \{D(\boldsymbol{\delta}, \boldsymbol{\beta}, g)\}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\mathbf{U}_{\boldsymbol{\beta},i}(\boldsymbol{\delta}, \boldsymbol{\beta}, g) - G(\boldsymbol{\delta}, \boldsymbol{\beta}, g)\mathbf{U}_{\boldsymbol{\gamma},i}(\boldsymbol{\delta}, \boldsymbol{\beta}, g)] \\ -\{D(\boldsymbol{\delta}, \boldsymbol{\beta}, g)\}^{-1}[E\{\partial \mathbf{U}_{\boldsymbol{\beta},i}(\boldsymbol{\delta}, \boldsymbol{\beta}, g)/\partial \boldsymbol{\delta}^\top\} \\ -G(\boldsymbol{\delta}, \boldsymbol{\beta}, g) E\{\partial \mathbf{U}_{\boldsymbol{\gamma},i}(\boldsymbol{\delta}, \boldsymbol{\beta}, g)/\partial \boldsymbol{\delta}^\top\}] A^{-1}(\boldsymbol{\delta}) \frac{1}{\sqrt{n}} \mathbf{R}(\boldsymbol{\delta}). \end{aligned}$$

We use D_0 and Σ_0 to denote the values of $D(\boldsymbol{\delta}, \boldsymbol{\beta}, g)$ and $\Sigma(\boldsymbol{\delta}, \boldsymbol{\beta}, g)$ at $\boldsymbol{\delta}_0, \boldsymbol{\beta}_0, g_0$, respectively. Theorem 4 describes the asymptotic properties of the proposed estimators.

Theorem 4 *Under identifiable observed likelihood* (2) *and conditions (B)-(J), if* $nh^2/\log(1/h) \to \infty$ *and* $nh^4 \to 0$, *we have*

(i) $\|\hat{g}_k - g_{0k}\| = O_p\{(N_n/n)^{1/2}\}$, $1 \leq k \leq d_2$,

(ii) $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N(0, D_0^{-1}\Sigma_0 D_0^{-1^\top})$,

where $\|\hat{g}_k - g_{0k}\|^2 = E\{\hat{g}_k(Z_k) - g_{0k}(Z_k)\}^2$.

The proof of Theorem 4 is provided in the Supplementary Material. In practical applications, the covariance matrix can be estimated using $\hat{D}^{-1}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g})\hat{\Sigma}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g})$ $\{\hat{D}^{-1}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g})\}^\top$, where

$$\hat{D}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g}) = n^{-1} \sum_{i=1}^{n} \{\partial \mathbf{U}_{\boldsymbol{\beta},i}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g})/\partial \boldsymbol{\beta}^\top - \hat{G}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g}) \partial \mathbf{U}_{\boldsymbol{\gamma},i}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g})/\partial \boldsymbol{\beta}^\top\},$$

$$\hat{G}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g}) = n^{-1} \sum_{i=1}^{n} \partial \mathbf{U}_{\boldsymbol{\beta},i}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g})/\partial \boldsymbol{\gamma}^\top \Big[\sum_{i=1}^{n} \partial \mathbf{U}_{\boldsymbol{\gamma},i}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g})/\partial \boldsymbol{\gamma}^\top\Big]^{-1},$$

$$\hat{\Sigma}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g}) = n^{-1} \sum_{i=1}^{n} \{m_i(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g}) m_i^\top(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g})\},$$

$$
\begin{aligned}
m_i(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g}) = \{\hat{D}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g})\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathbf{U}_{\boldsymbol{\beta},i}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g}) - \hat{G}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g})\mathbf{U}_{\boldsymbol{\gamma},i}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g})] \\
-\{\hat{D}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g})\}\Big[\frac{1}{n}\sum_{i=1}^{n}\partial\mathbf{U}_{\boldsymbol{\beta},i}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g})/\partial\boldsymbol{\delta}^{\top} \\
-\hat{G}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g})\frac{1}{n}\sum_{i=1}^{n}\partial\mathbf{U}_{\boldsymbol{\gamma},i}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g})/\partial\boldsymbol{\delta}^{\top}]\hat{A}^{-1}(\hat{\boldsymbol{\delta}})\frac{1}{\sqrt{n}}\hat{\mathbf{U}}(\hat{\boldsymbol{\delta}}).
\end{aligned}
$$

4 Simulations

In this section, we present the simulation results of the proposed estimators introduced in Sect. 3. We consider three types of response models: Logistic regression, quasi-Poisson regression, and truncated normal regression. These models are subject to different missing data mechanisms, namely logistic, probit, and complementary log-log models.

The covariate vector $(\mathbf{X}^{\top}, \mathbf{Z}^{\top})^{\top} = (X_1, X_2, Z_1, Z_2)^{\top}$, and $g(\mathbf{Z}) = g_1(Z_1) + g_2(Z_2) = \sin(4\pi Z_1) + 5(Z_2 - 0.5)^2 - 5/12$, where Z_1 and Z_2 are independently uniformly distributed on [0, 1]. We assume that T_1 and T_2 are normally distributed in $\mathbb{R}^2$ with

$$
\mu = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.
$$

To account for the dependence between $\mathbf{X}$ and $\mathbf{Z}$, we assume the following relationship: $X_1 = T_1 + 0.5(Z_1 + Z_2)$ and $X_2 = T_2 + 0.5(Z_1 + Z_2)$.

The simulation models are designed as follows:

Binary case: The response variable Y follows a Bernoulli distribution

$$
P(Y = 1|\mathbf{X}, \mathbf{Z}) = \text{expit}\{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + g_1(Z_1) + g_2(Z_2)\},
$$

where $(\beta_0, \beta_1, \beta_2)^{\top} = (-1, 1, -1)^{\top}$. The missing data mechanism model is of a logistic form

$$
\pi(Y, \mathbf{X}, \mathbf{Z}; \alpha, \boldsymbol{\theta}) = \text{expit}(\alpha Y + \theta_0 + \theta_1 X_1 + \theta_2 Z_1),
$$

with $(\alpha, \theta_0, \theta_1, \theta_2)^{\top} = (0.2, 1.8, 0.2, -0.2)^{\top}$, or a probit form

$$
\pi(Y, \mathbf{X}, \mathbf{Z}; \alpha, \boldsymbol{\theta}) = \Phi(\alpha Y + \theta_0 + \theta_1 X_1 + \theta_2 Z_1),
$$

with $(\alpha, \theta_0, \theta_1, \theta_2)^{\top} = (0.5, 1.3, -0.2, 0.2)^{\top}$, or a complementary log-log form

$$
\pi(Y, \mathbf{X}, \mathbf{Z}; \alpha, \boldsymbol{\theta}) = 1 - \exp\{-\exp(\alpha Y + \theta_0 + \theta_1 X_1 + \theta_2 Z_1)\},
$$

with $(\alpha, \theta_0, \theta_1, \theta_2)^\top = (0.3, 0.9, -0.2, 0.2)^\top$. The first case leads to that the percentage of complete data is about 90.1%, the second one is about 90.8% and the third one is about 89.4%.
Quasi-Poisson case: Y follows quasi-Possion with conditional expectation

$$P(Y = 0|\mathbf{X}, \mathbf{Z}) = 1/\exp\{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + g_1(Z_1) + g_2(Z_2)\},$$

where $(\beta_0, \beta_1, \beta_2)^\top = (0.5, -0.5, 0.5)^\top$ and dispersion parameter $\phi = 1.5$. The missing data mechanism model is of a logistic form

$$\pi(Y, \mathbf{X}, \mathbf{Z}; \alpha, \boldsymbol{\theta}) = \text{expit}(\alpha Y + \theta_0 + \theta_1 X_1 + \theta_2 Z_1),$$

with $(\alpha, \theta_0, \theta_1, \theta_2)^\top = (0.3, 1.5, 0.2, 0.2)^\top$, or a probit form

$$\pi(Y, \mathbf{X}, \mathbf{Z}; \alpha, \boldsymbol{\theta}) = \Phi(\alpha Y + \theta_0 + \theta_1 X_1 + \theta_2 Z_1),$$

with $(\alpha, \theta_0, \theta_1, \theta_2)^\top = (0.15, 0.8, 0.2, -0.2)^\top$, or a complementary log-log form

$$\pi(Y, \mathbf{X}, \mathbf{Z}; \alpha, \boldsymbol{\theta}) = 1 - \exp\{-\exp(\alpha Y + \theta_0 + \theta_1 X_1 + \theta_2 Z_1)\},$$

with $(\alpha, \theta_0, \theta_1, \theta_2)^\top = (0.3, 0.8, -0.1, -0.1)^\top$. The first case leads to that the percentage of complete data is about 90.2%, the second one is about 88.7% and the third one is about 88.9%.
Truncated normal case: Y is generated according to the following model:

$$Y \sim TN(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + g_1(Z_1) + g_2(Z_2), \sigma^2, \mu(\mathbf{X}, \mathbf{Z}) - c, \mu(\mathbf{X}, \mathbf{Z}) + c),$$

where $\mu(\mathbf{X}, \mathbf{Z}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + g_1(Z_1) + g_2(Z_2)$. The parameter vector is set at $(\beta_0, \beta_1, \beta_2, \sigma^2, c)^\top = (1, 2, 2, 1, 2)^\top$. The indicator variable r is generated from Bernoulli distribution with probability function being specified as a logistic form

$$\pi(Y, \mathbf{X}, \mathbf{Z}; \alpha, \boldsymbol{\theta}) = \text{expit}(\alpha Y + \theta_0 + \theta_1 X_1 + \theta_2 Z_1),$$

with $(\alpha, \theta_0, \theta_1, \theta_2)^\top = (-0.5, 2, 1.5, 0.5)^\top$, or a probit form

$$\pi(Y, \mathbf{X}, \mathbf{Z}; \alpha, \boldsymbol{\theta}) = \Phi(\alpha Y + \theta_0 + \theta_1 X_1 + \theta_2 Z_1),$$

with $(\alpha, \theta_0, \theta_1, \theta_2)^\top = (-0.2, 1, 1, -0.6)^\top$, or a complementary log-log form

$$\pi(Y, \mathbf{X}, \mathbf{Z}; \alpha, \boldsymbol{\theta}) = 1 - \exp\{-\exp(\alpha Y + \theta_0 + \theta_1 X_1 + \theta_2 Z_1)\},$$

with $(\alpha, \theta_0, \theta_1, \theta_2)^\top = (-0.2, 1, 0.7, -0.5)^\top$. The first case leads to that the percentage of complete data is about 82.6%, the second one is about 85.7% and the third one is about 86.3%.

The simulation study was conducted with different sample sizes: $n = 2000$ for the Binary case, $n = 1000$ for the quasi-Poisson case, and $n = 500$ for the truncated normal case. The reason for conducting simulations with three different sample sizes is rooted in the fact that when the population distribution is highly imbalanced, such as in a binary scenario, might need a substantial sample size for the central limit theorem to kick in and produce sampling distributions that approximate a normal distribution. Ideally, Binary and skewed multi-category discrete predictors demand larger sample sizes compared to normally-distributed continuous predictors (Olvera Astivia et al. 2019).

The number of knots N_n was determined automatically using the R package mgcv. The proposed estimators were implemented in R using the iteration algorithm described in Sect. 3. The simulation results based on 1000 runs are summarized in Tables 1, 2, and 3. These tables present the bias, standard deviation (SD), approximate 95% confidence intervals (CI), and coverage rate (CR) of the estimated parameters. The confidence intervals were constructed using the formula "estimator ± 1.96SE," where SE is the square root of the diagonal elements of the matrix $\hat{A}^{-1}(\hat{\boldsymbol{\delta}})\hat{B}(\hat{\boldsymbol{\delta}})\hat{A}^{-1}(\hat{\boldsymbol{\delta}})^{\top}$ and $\hat{D}^{-1}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g})\hat{\Sigma}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g})\hat{D}^{-1}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}, \hat{g})^{\top}$. Figures 1, 2, and 3 depict the mean of the fitted nonparametric functions and the approximate 95% confidence bands (CB). Overall, in the three examples with different missing data mechanism models, both the parameter estimators and the nonparametric function estimators perform well. As we might expect, the estimators in the same response model with different missing data mechanisms show the similar bias and variance because of the similar missing rate. While the estimators in the truncated normal case exhibit the highest bias and variance due to the smallest sample size and highest missing rate, compared to the binary and quasi-Poisson cases. For Brnary case under the missingness mechanism of logistic form, the estimate θ_2 has a slightly lower coverage rate.

In the study conducted by Du et al. (2023), the analysis involves utilizing data with nonresponse alongside a parametric distribution $p(y|\mathbf{x}, \mathbf{z})$ that is a member of the exponential family. The optimal estimator can be obtained by maximizing the observed likelihood. In this paper, we consider nonparametric $p(y|\mathbf{x}, \mathbf{z})$ and constructs estimating equations for mean of response based on the inverse probability weighting. Thus, our method expands the scope of applicability for these models. And with large samples, even if $p(y|\mathbf{x}, \mathbf{z})$ belongs to the exponential family, the difference between the two methods almost disappears.

In order to assess the stability of the proposed inference method, we consider scenarios where the missingness type or the missingness mechanism model is misspecified. We assume that the response variable Y is generated according to

$$Y \sim N(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + g_1(Z_1) + g_2(Z_2), \sigma^2),$$

the parameter vector is set at $(\beta_0, \beta_1, \beta_2, \sigma^2)^{\top} = (1, 2, 2, 1)^{\top}$. The indicator variable r is generated from Bernoulli distribution with probability function being specified as a logistic form

$$\pi(Y, \mathbf{X}, \mathbf{Z}; \alpha, \boldsymbol{\theta}) = \text{expit}(\alpha Y + \theta_0 + \theta_1 X_1 + \theta_2 Z_1),$$

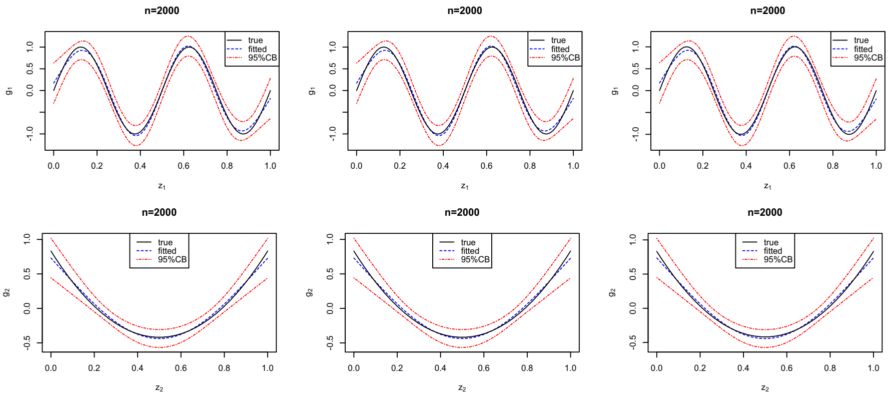

Fig. 1 The left graph is for Logistic, the middle graph is for Probit and the right graph is for Clog-log

with $(\alpha, \theta_0, \theta_1, \theta_2)^\top = (-0.5, 2, 1.5, 0.5)^\top$, and The percentage of complete data in the dataset is approximately 82.4%. Here we mainly focus on that the missingness type is mis-specified to be missing completely at random or missing at random, and the missingness mechanism is mis-specified to be of the probit form. The sample size is 500, and based on 1000 simulation runs, Table 4 presents the bias, standard deviation (SD), approximately 95% confidence intervals (CI) of the parameters with coverage rate (CR). Figure 4 shows the mean of the fitted nonparametric functions and the approximately 95% confidence band (CB). The performance of the proposed method is not good when using complete data or missing at random mechanism. One of them fails completely. However, when the missingness mechanism is misspecified as a probit form, the estimators of both parameters and nonparametric functions are less affected. The reason is that the performance with the probit and logistic model is very similar, and in this case, misspecification of the response model is not a serious problem (Morikawa and Kim 2021). Additionally, we observed that the probit and logistic models yielded almost identical outcomes across the three types of response models.

Estimation is not possible when the parameter is non-identifiable. Despite providing simulation results, the estimators are challenging to compute due to fluctuations. The missing mechanism is assumed to follow a logistic form, that is $h(\alpha y+\theta_0+\theta_1 z_1) = \log\{\text{expit}(\alpha y+\theta_0+\theta_1 z_1)\}$, and $p(y|\mathbf{x}, \mathbf{z}; \boldsymbol{\beta}, g, \phi) = \exp[-y/\{g_1(z_1)+\beta_0\}]/\{g_1(z_1)+\beta_0\}$, and $g_1(z_1) + \beta_0 > 0$. The condition given in (3) reduces to

$$\begin{aligned}&\log\{\text{expit}(\alpha y + \theta_0 + \theta_1 z_1)\} - y/\{g_1(z_1) + \beta_0\} - \log\{g_1(z_1) + \beta_0\}\\ &\quad = \log\{\text{expit}(\alpha^* y + \theta_0^* + \theta_1^* z_1)\} - y/\{g_1^*(z_1) + \beta_0^*\} - \log\{g_1^*(z_1) + \beta_0^*\}.\end{aligned}$$

For example, we can take that

$$\begin{aligned}(\alpha, \theta_0, \theta_1, \beta_0, g_1(z_1))^\top &= (1, -1, -1, 1 - e^{-1}, e^{-1} - e^{-z_1-1})^\top,\\ (\alpha^*, \theta_0^*, \theta_1^*, \beta_0^*, g_1^*(z_1))^\top &= (-1, 1, 1, e - 1, e^{z_1+1} - e)^\top,\end{aligned}$$

Table 1 Bias $\times 10^2$, standard deviation (SD) $\times 10^2$, confidence interval (CI) and coverage rate (CR) of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ using three link functions for Binary case

	Logistic		Probit		Cloglog	
Variable	Bias(SD)	CI(CR)	Bias(SD)	CI(CR)	Bias(SD)	CI(CR)
β_0	0.61 (11.7)	[−1.241, −0.746] (0.960)	0.05 (11.8)	[−1.304, −0.695] (0.967)	1.04 (11.7)	[−1.306, −0.673] (0.988)
β_1	0.26 (8.5)	[0.833, 1.172] (0.952)	0.49 (8.9)	[0.812, 1.198] (0.958)	0.27 (8.6)	[0.804, 1.191] (0.968)
β_2	−0.30(6.6)	[−1.138, −0.868] (0.967)	−0.19 (6.8)	[−1.145, −0.859] (0.955)	−0.01(6.7)	[−1.139, −0.861] (0.967)
α	−0.29 (7.1)	[0.339, 0.655] (0.956)	−0.23 (2.5)	[0.326, 0.669] (0.999)	−0.06 (4.2)	[0.123, 0.476] (0.996)
θ_0	1.24 (10.2)	[1.630, 1.995] (0.944)	−0.85 (9.9)	[1.100, 1.483] (0.928)	1.28(5.2)	[0.727, 1.098] (0.995)
θ_1	−1.15(7.3)	[0.035, 0.342] (0.924)	0.66 (6.2)	[−0.354, −0.033] (0.938)	−0.92 (3.7)	[−0.367, −0.052] (1.000)
θ_2	1.31(10.1)	[−0.341, −0.033](0.897)	−0.05(5.4)	[0.039, 0.360] (0.986)	0.96(4.1)	[0.051, 0.368] (0.996)

Table 2 Bias $\times 10^2$, standard deviation (SD) $\times 10^2$, confidence interval (CI) and coverage rate (CR) of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ using three link functions for quasi-Possion case

	Logistic		Probit		Cloglog	
Variable	Bias(SD)	CI(CR)	Bias(SD)	CI(CR)	Bias(SD)	CI(CR)
β_0	1.05 (6.3)	[0.363, 0.666] (0.963)	1.78 (7.3)	[0.374, 0.662] (0.935)	2.82 (6.8)	[0.411, 0.640] (0.928)
β_1	−0.10 (3.8)	[−0.629, −0.374] (0.981)	−0.11 (4.6)	[−0.607, −0.395] (0.973)	−0.31 (4.6)	[−0.582, −0.417] (0.958)
β_2	0.17 (2.8)	[0.384, 0.618] (0.984)	0.20 (3.3)	[0.414, 0.590] (0.980)	0.03(3.2)	[0.432, 0.568] (0.977)
α	1.63(12.7)	[0.111, 0.521] (0.908)	1.50 (8.2)	[0.051, 0.279] (0.905)	0.40 (5.2)	[0.096, 0.312] (0.941)
θ_0	−2.41 (24.1)	[0.635, 2.316] (0.960)	−1.57 (10.1)	[0.598, 0.971] (0.928)	1.91 (8.0)	[0.596, 1.042] (0.980)
θ_1	1.19 (9.7)	[−0.110, 0.534] (0.973)	1.52(5.4)	[0.053, 0.378] (0.986)	−1.59 (4.5)	[−0.363, −0.069] (0.995)
θ_2	1.12(17.7)	[−0.562, 0.985] (0.971)	−1.47(10.6)	[−0.444, 0.014] (0.946)	1.76(6.7)	[0.016, 0.419] (0.980)

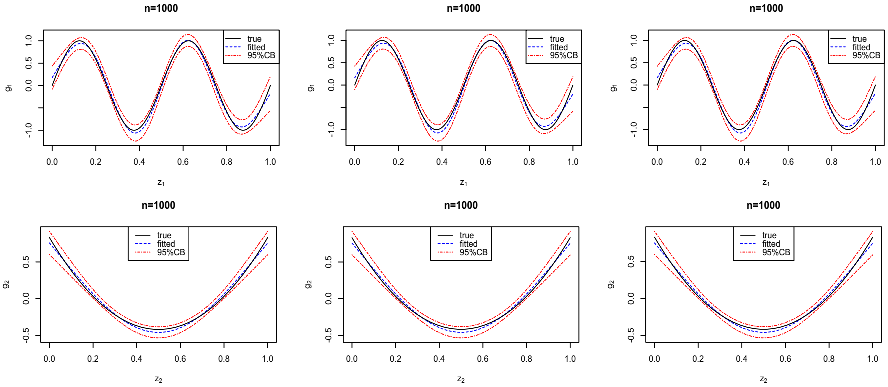

Fig. 2 The left graph is for Logistic, the middle graph is for Probit and the right graph is for Clog-log

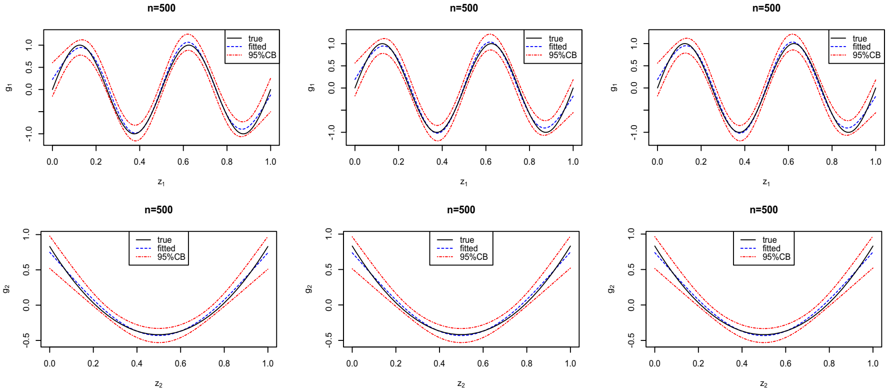

Fig. 3 The left graph is for Logistic, the middle graph is for Probit and the right graph is for Clog-log

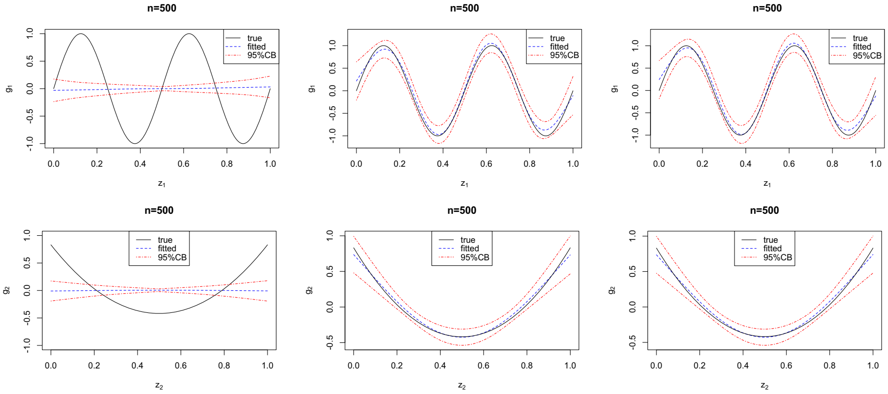

Fig. 4 The left graph is based on complete data, the middle graph on missing at random and the right graph on the probit missing data mechanism

Table 3 Bias $\times 10^2$, standard deviation (SD) $\times 10^2$, confidence interval (CI) and coverage rate (CR) of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ using three link functions for truncated normal case

	Logistic		Probit		Cloglog	
Variable	Bias(SD)	CI(CR)	Bias(SD)	CI(CR)	Bias(SD)	CI(CR)
β_0	−5.50(9.0)	[0.728, 1.162](0.920)	−2.51(8.9)	[0.741, 1.209] (0.940)	−2.78(8.9)	[0.734, 1.210] (0.948)
β_1	0.35(6.2)	[1.848, 2.159](0.969)	0.11 (6.1)	[1.859, 2.143] (0.950)	0.23(6.1)	[1.859, 2.146] (0.960)
β_2	−0.63(4.4)	[1.865, 2.122](0.957)	−0.02(4.3)	[1.899, 2.100] (0.959)	−0.03(4.3)	[1.898, 2.101] (0.963)
α	0.73(4.7)	[−0.626, −0.360](0.982)	0.54 (1.8)	[−0.260, −0.129] (0.972)	−0.19 (1.5)	[−0.267, −0.137] (0.972)
θ_0	4.56(28.6)	[0.761, 3.330](0.980)	−1.93 (13.4)	[0.515, 1.446] (0.995)	3.14(12.6)	[0.611, 1.452] (0.984)
θ_1	−6.02(21.7)	[0.656, 2.223] (0.990)	0.05 (9.9)	[0.719, 1.282] (1.000)	−1.48 (9.7)	[0.409, 0.961](0.998)
θ_2	1.33(16.2)	[0.002, 1.024] (0.995)	−0.97(7.7)	[−0.814, −0.405] (0.983)	1.72(7.2)	[−0.663, −0.303](0.963)

Table 4 Bias $\times 10^2$, standard deviation (SD) $\times 10^2$, confidence interval (CI) and coverage rate (CR) of $\boldsymbol{\beta}$ and ϕ for misspecification

	Mising completely at random		Mising at random		Mis-specified to be Probit	
Variable	Bias(SD)	CI(CR)	Bias(SD)	CI(CR)	Bias(SD)	CI(CR)
β_0	−109.27(10.0)	[−0.312, 0.131](0.000)	−14.20(11.4))	[0.594, 1.123] (0.804)	−5.57(11.5)	[0.644, 1.245] (0.929)
β_1	−197.61(6.7)	[−0.117, 0.165](0.000)	2.40 (7.2)	[1.813, 2.235] (0.937)	0.41(7.5)	[1.827, 2.181] (0.952)
β_2	−204.96(5.2)	[−0.158, 0.059](0.000)	−4.44(5.1)	[1.766, 2.145] (0.905)	−0.63(5.7)	[1.870, 2.117] (0.950)

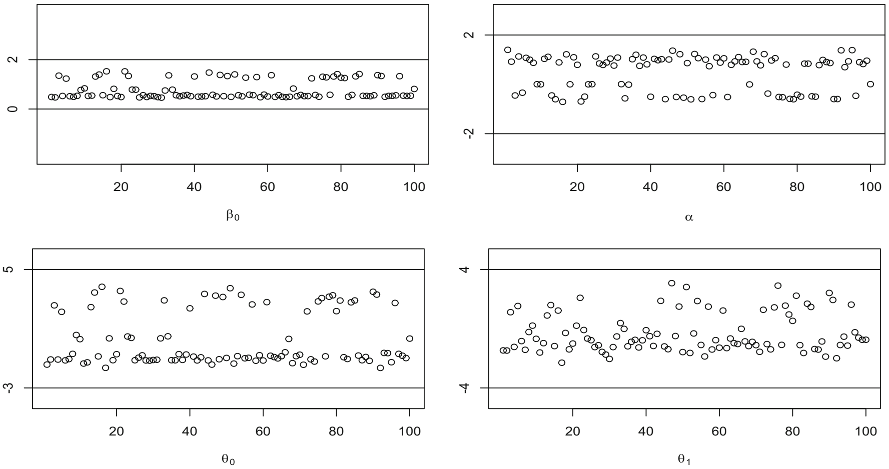

Fig. 5 The estimators of β_0, α, θ_0 and θ_1

which satisfies the above formula. Since

$$(\alpha, \theta_0, \theta_1, \beta_0, g_1(z))^\top \neq (\alpha^*, \theta_0^*, \theta_1^*, \beta_0^*, g_1^*(z))^\top .$$

Hence, this model is considered non-identifiable. We generated data from the non-identifiable model with a sample size of 1000. Based on 100 simulation runs, Figs. 5 illustrates the estimators of β_0, α, θ_0, and θ_1 varied between two sets of values.

5 Real data analysis

The CHIP survey (2013) aims to measure the distribution of personal income and related economic factors in rural, migrant, and urban areas of China (Sicular et al. 2020). The survey includes data from cities and towns in fifteen provinces, which are representative of different regions in the country. These provinces include Liaoning, Shanxi, Jiangsu, Shandong, Guangdong, Anhui, Henan, Sichuan, Hunan, Hubei, Gansu, Xinjiang, Yunnan, Beijing, and Chongqing. The selected provinces represent the north, eastern coastal areas, interior regions, and western regions of China.

In this study, the analysis focuses on urban data, which consists of a sample of 12,233 individuals. The percentage of missingness in the data is 22.4%. Instead of assuming a linear relationship between work experience and the log of income, a smooth function is used, similar to the Mincer earnings function. The model is specified as follows:

$$\log \mathrm{E} = \beta_0 + \beta_1 \mathrm{S} + g_1(\mathrm{Exper}) + \varepsilon, \tag{14}$$

In this model, the logarithm of earnings $\log \mathrm{E}$ is related to years of schooling S and work experience (Exper), which is calculated as age $-$ S $-$ 6. The relationship is subject to an unobserved random error (ε) with variance ϕ. Without considering the cost of

Table 5 Estimate and standard deviation (SD) for the parameters of (14) under the nonignorable missingness (NIM) and Missing at random (MAR)

	Nonignorable missingness			Missing at random		
Variable	Estimates	SD	t-value	Estimates	SD	t-value
β_0	9.3792	0.0325	288.717	9.1882	0.0347	264.740
β_1	0.0830	0.0027	29.800	0.0967	0.0029	32.991
α	−0.2410	0.0112	−21.586	–	–	–
θ_0	2.7663	0.3058	9.045	1.9479	0.3087	6.311
θ_1	1.5436	0.0752	20.532	1.1319	0.0764	13.853
θ_2	−0.8684	0.0585	−14.835	−1.0882	0.0553	−18.5059

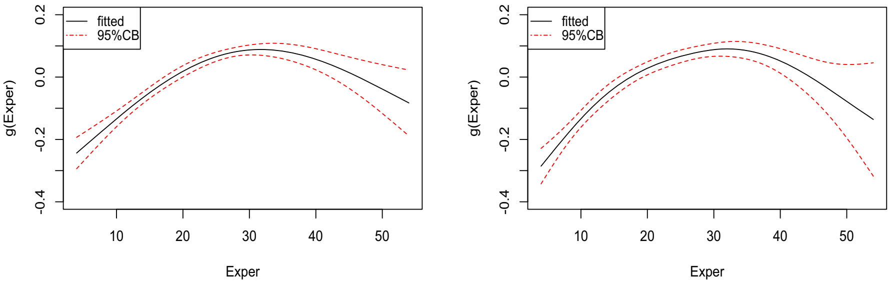

Fig. 6 The left graph is obtained under the nonignorable missingness and the right graph is under missing at random

education, the rates of return to schooling can be calculated as

$$\partial \log \mathrm{E}/\partial \mathrm{S} = \beta_1.$$

The missing data mechanism is modeled using the following model:

$$\begin{aligned} &P(r = 1|\log \mathrm{E}, \log \mathrm{S}, \log \mathrm{Exper}) \\ &\quad = \operatorname{expit}(\alpha \log \mathrm{E} + \theta_0 + \theta_1 \log \mathrm{S} + \theta_2 \log \mathrm{Exper}). \end{aligned} \tag{15}$$

To account for the large values and fluctuations in schooling and experience, we replace S and Exper with log S and log Exper. log S represents the logarithm of 1+ years of education because uneducated groups exist, and log Exper represents the logarithm of work experience.

Table 5 presents parameter estimates for models (14) and (15) under the nonignorable missingness and missing at random. The results show that under the nonignorable missing mechanism, log-income has a significant negative effect on the probability of missingness. Moreover, the rates of return to schooling in China is 8.30% under the nonignorable missingness assumption and 9.67% under the assumption of missingness at random. These estimators are consistent with the existing literature (Gao and Smyth 2015; Kang and Peng 2012), which suggests that the returns to education in China range from 8% to 10%. It is worth noting that the rates of return to schooling in China have stagnated or even declined after 2005 due to factors such as education expansion and labor mobility (Cai and Wang 2010). The rates of return to schooling in China under the nonignorable missingness is smaller than that under missing at random, suggesting that it is required to model the missingness mechanism. Figure 5 illustrates the functional relationship between log-income and work experience, revealing an inverted U-shaped pattern. This finding is consistent with the classic hypothesis of the Mincerian earnings equation, which suggests that there is an optimal level of work experience that maximizes income.

6 Conclusions

In this study, semiparametric estimators have been developed specifically for handling nonignorable missing data. These estimators are designed to accommodate the logistic model, probit model, and complementary log-log model, which are commonly used to characterize the missing data mechanism. The instrumental variable assumption ensures identifiability without requiring additional assumptions, giving it an advantage over the assumptions proposed by Tchetgen Tchetgen and Wirth (2017) and Sun et al. (2018). If the score function $S_\mu\{\mathbf{X}, \mathbf{Z}; \mu, \upsilon(\cdot)\}$ can be written as $a\{\mathbf{X}, \mathbf{Z}; \mu, \upsilon(\cdot)\}Y + b\{\mathbf{X}, \mathbf{Z}; \mu, \upsilon(\cdot)\}$, such as in the case of samples following exponential family, identifiability can be achieved without relying on instrumental variables. By employing the kernel method and spline method, we extend the generalized linear regression model to the generalized additive partial linear model when the distribution of the response variable is unknown.

The approaches presented in Cui and Zhou (2017) and Morikawa et al. (2017) for estimating the missing mechanism model, due to their lack of dimensionality reduction, become invalid when dealing with numerous covariates. In this paper, we utilize dimension reduction techniques to achieve readily achievable univariate kernel estimation. While univariate kernel estimation may compromise estimator efficiency, it significantly reduces computational complexity. To enhance estimation efficiency, one can employ the estimation method based on the effective score introduced by Morikawa and Kim (2021). Nevertheless, under the assumption that $p(y|\mathbf{x}, \mathbf{z})$ belongs to the exponential family, the optimal estimator can be derived by maximizing the observed likelihood (Morikawa and Kim 2021).

There are many directions worthy of further research. A possible extension in this research area involves transforming the identifiability of the observation likelihood into the identifiability of the parameters of interest, such as mean functionals (Li et al. 2021). Indeed, the development of doubly robust estimation methods and efficient estimation techniques for nonignorable missing data is a crucial research area. Furthermore, incorporating more sophisticated structures in the missing mechanism model is another promising research direction.

Supplementary Information

Supplementary material is available online at Statistical Papers.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00362-023-01522-0.

Acknowledgements We thank the Editor, Associate Editor and referees, as well as our financial sponsors. This work was supported by the National Natural Science Foundation of China (Grant No. 11871173), and the National Statistical Science Research Project (Grant No. 2020LZ09).

References

An L, Brown DG (2008) Survival analysis in land change science: Integrating with GIScience to address temporal complexities. Ann Assoc Am Geogr 98(2):323–344. https://doi.org/10.1080/00045600701879045

Beppu K, Morikawa K, Im J (2022) Imputation with verifiable identification condition for nonignorable missing outcomes. arXiv:2204.10508

Cai F, Wang M (2010) Growth and structural changes in employment in transition china. J Comp Econ 38(1):71–81. https://doi.org/10.1016/j.jce.2009.10.006

Cameron A, Trivedi P (2005) Microeconometrics: methods and applications. Cambridge University Press, Cambridge

Chen J, Shao J, Fang F (2021) Instrument search in pseudo-likelihood approach for nonignorable nonresponse. Ann Inst Stat Math 73:519–533. https://doi.org/10.1007/s10463-020-00758-z

Cui X, Zhou Y (2017) Estimated conditional score function for missing mechanism model with nonignorable nonresponse. Sci China Math 60(7):1197–1218. https://doi.org/10.1007/s11425-015-9014-1

Cui X, Guo J, Yang G (2017) On the identifiability and estimation of generalized linear models with parametric nonignorable missing data mechanism. Comput Stat Data Anal 107:64–80. https://doi.org/10.1016/j.csda.2016.10.017

Cui L-E, Zhao P, Tang N (2022) Generalized empirical likelihood for nonsmooth estimating equations with missing data. J Multivar Anal 190:104907. https://doi.org/10.1016/j.jmva.2021.104907

Du J, Li Y, Cui X (2023) Identification and estimation of generalized additive partial linear models with nonignorable missing response. Commun Math Stat. https://doi.org/10.1007/s40304-022-00284-9

Fan J, Gijbels I, Hu T-C, Huang L-S (1996) A study of variable bandwidth selection for local polynomial regression. Stat Sin 6:113–127

Fang F, Shao J (2016) Model selection with nonignorable nonresponse. Biometrika 103:861–874. https://doi.org/10.1093/biomet/asw039

Gao W, Smyth R (2015) Education expansion and returns to schooling in urban china, 2001–2010: evidence from three waves of the china urban labor survey. J Asia Pac Econ 20(2):178–201. https://doi.org/10.1080/13547860.2014.970607

Hamori S, Motegi K, Zhang Z (2019) Calibration estimation of semiparametric copula models with data missing at random. J Multivar Anal 173:85–109. https://doi.org/10.1016/j.jmva.2019.02.003

Hattaway JT (2010) Parameter estimation and hypothesis testing for the truncated normal distribution with applications to introductory statistics grades. Brigham Young University, Provo

Kang L, Peng F (2012) Real wage cyclicality in urban china. Econ Lett 115(2):141–143. https://doi.org/10.1016/j.econlet.2011.12.009

Kott PS, Chang T (2010) Using calibration weighting to adjust for nonignorable unit nonresponse. J Am Stat Assoc 105(491):1265–1275. https://doi.org/10.1198/jasa.2010.tm09016

Krosnick JA, Holbrook AL, Berent MK, Carson RT, Michael Hanemann W, Kopp RJ, Cameron Mitchell R, Presser S, Ruud PA, Kerry Smith V et al (2002) The impact of" no opinion" response options on data quality: non-attitude reduction or an invitation to satisfice? Public Opin Q 66(3):371–403. https://doi.org/10.1086/341394

Li W, Miao W, Tchetgen Tchetgen E (2021) Nonparametric inference about mean functionals of nonignorable nonresponse data without identifying the joint distribution. J R Stat Soc Ser B. https://doi.org/10.1093/jrsssb/qkad047

Li M, Ma Y, Zhao J (2022) Efficient estimation in a partially specified nonignorable propensity score model. Comput Stat Data Anal 174:107322. https://doi.org/10.1016/j.csda.2021.107322

Little RJ, Rubin DB (2019) Statistical analysis with missing data. Wiley, Hoboken

Miao W, Tchetgen Tchetgen EJ (2016) On varieties of doubly robust estimators under missingness not at random with a shadow variable. Biometrika 103(2):475–482. https://doi.org/10.1093/biomet/asw016

Miao W, Ding P, Geng Z (2016) Identifiability of normal and normal mixture models with nonignorable missing data. J Am Stat Assoc 111(516):1673–1683

Morikawa K, Kim JK (2021) Semiparametric optimal estimation with nonignorable nonresponse data. Ann Stat 49(5):2991–3014. https://doi.org/10.1214/21-AOS2070

Morikawa K, Kim JK, Kano Y (2017) Semiparametric maximum likelihood estimation with data missing not at random. Can J Stat 45(4):393–409

Nabi R, Bhattacharya R (2022) On testability and goodness of fit tests in missing data models. Uncertain Artif Intell. https://doi.org/10.48550/arXiv.2203.00132

Olvera Astivia OL, Gadermann A, Guhn M (2019) The relationship between statistical power and predictor distribution in multilevel logistic regression: a simulation-based approach. BMC Med Res Methodol 19(1):1–20. https://doi.org/10.1186/s12874-019-0742-8

Shao J, Wang L (2016) Semiparametric inverse propensity weighting for nonignorable missing data. Biometrika 103(1):175–187

Shao Y, Wang L (2022) Generalized partial linear models with nonignorable dropouts. Metrika 85(2):223–252. https://doi.org/10.1007/s00184-021-00828-z

Sicular T, Li S, Yue X, Sato H (2020) Changing trends in China's inequality: evidence, analysis, and prospects. Oxford University Press, New York

Sun B, Liu L, Miao W, Wirth K, Robins J, Tchetgen EJT (2018) Semiparametric estimation with data missing not at random using an instrumental variable. Stat Sin 28(4):1965. https://doi.org/10.5705/ss.202016.0324

Tang N, Zhao P, Zhu H (2014) Empirical likelihood for estimating equations with nonignorably missing data. Stat Sin 24(2):723–747. https://doi.org/10.5705/ss.2012.254

Tchetgen Tchetgen EJ, Wirth KE (2017) A general instrumental variable framework for regression analysis with outcome missing not at random. Biometrics 73(4):1123–1131. https://doi.org/10.1111/biom.12670

Tsiatis AA (2006) Semiparametric theory and missing data. Springer, New York

Wang L, Yang L (2007) Spline-backfitted kernel smoothing of nonlinear additive autoregression model. Ann Stat 35(6):2474–2503. https://doi.org/10.1214/009053607000000488

Wang L, Liu X, Liang H, Carroll RJ (2011) Estimation and variable selection for generalized additive partial linear models. Ann Stat 39(4):1827. https://doi.org/10.1214/11-AOS885

Wang S, Shao J, Kim JK (2014) An instrumental variable approach for identification and estimation with nonignorable nonresponse. Stat Sin 24(2):1097–1116. https://doi.org/10.5705/ss.2012.074

Wang L, Shao J, Fang F (2021) Propensity model selection with nonignorable nonresponse and instrument variable. Stat Sin 31(2):647–672

Xue L, Yang L (2006) Additive coefficient modeling via polynomial spline. Stat Sin 16:1423–1446

Zhao J, Ma Y (2022) A versatile estimation procedure without estimating the nonignorable missingness mechanism. J Am Stat Assoc 117(540):1916–1930. https://doi.org/10.1080/01621459.2021.1893176

Zhao J, Shao J (2015) Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. J Am Stat Assoc 110(512):1577–1590. https://doi.org/10.1080/01621459.2014.983234

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.