Computational Statistics and Data Analysis 79 (2014) 91–112

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Empirical likelihood inference in linear regression with nonignorable missing response

Cuizhen Niu[a], Xu Guo[b,c], Wangli Xu[a,*], Lixing Zhu[b]

[a] School of Statistics, Renmin University of China, Beijing, China

[b] Department of Mathematics, Hong Kong Baptist University, Hong Kong

[c] College of Economics and Management, Nanjing University of Aeronautics and Astronautics, China

ARTICLE INFO

Article history:
Received 24 September 2013
Received in revised form 25 February 2014
Accepted 6 May 2014
Available online 28 May 2014

Keywords:
Empirical likelihood inference
Nonignorable missing response
Linear regression
Inverse probability weighted
Imputed empirical likelihood

ABSTRACT

Parameter estimation for nonignorable nonresponse data is a challenging issue as the missing mechanism is unverified in practice and the parameters of response probabilities need to be estimated. This article aims at applying the empirical likelihood to construct the confidence intervals for the parameters of interest in linear regression models with nonignorable missing response data and the nonignorable missing mechanism is specified as an exponential tilting model. Three empirical likelihood ratio functions based on weighted empirical likelihood and imputed empirical likelihood are defined. It is proved that, except one that is chi-squared distributed, all the others are asymptotically weighted chi-squared distributed whenever the tilting parameter is either given or estimated. The asymptotic normality for the related parameter estimates is also investigated. Simulation studies are conducted to evaluate the finite sample performance of the proposed estimates in terms of coverage probabilities and average widths for the confidence intervals of parameters. A real data analysis is analyzed for illustration.

1. Introduction

Consider the classical linear regression model

$$y_i = x_i^\tau \beta + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where β is a $d \times 1$ vector of unknown regression parameter and ε_i's are independent and identically distributed (i.i.d.) random errors with conditional mean $E(\varepsilon|X) = 0$. Throughout this paper, we focus on the situation that some of the responses y_i in a sample of size n may be missing and all the covariates or auxiliary variables x_i's are observed completely. In this way, we obtain the following incomplete observations

$$(x_i, y_i, \delta_i), \quad i = 1, \ldots, n,$$

where δ_i is a missing indicator for ith individual and $\delta_i = 0$ if y_i is missing, otherwise $\delta_i = 1$. Often, the missing mechanism missing at random (MAR) is a common assumption for statistical analysis in the presence of missing data and is reasonable in many practical situations. Nevertheless, sometimes, there may be a concern that nonresponse is related to the value of the unobserved outcome variable y_i itself, even after controlling for x_i. The MAR mechanism then would become invalid. For

[*] Corresponding author.
E-mail address: wxu.stat@gmail.com (W. Xu).

http://dx.doi.org/10.1016/j.csda.2014.05.005
0167-9473/© 2014 Published by Elsevier B.V.

example, in the surveys about income or the history of committing, the nonresponse rates tend to be related to the values of nonresponses. Instead of the classical assumption MAR for missing data, the present paper assumes that missing response data are either nonignorable or not missing at random (NMAR).

Nonignorable missing is a ubiquitous existing problem in various disciplines, for example, medical research, clinical trials and longitudinal studies and in recent decades, there has been a number of literatures for the analysis of nonignorable missing values. Based on the exponential tilting model for the response probability, Kim and Yu (2011) proposed a semiparametric estimation method of mean functions with nonignorable missing data and derived the $\sqrt{n}$-consistency when the tilting parameter is either given or estimated. Zhao et al. (2013) applied the empirical likelihood to the inference of mean functionals with nonignorable missing response data when the inverse probability weighted methods with and without auxiliary information are used, and the asymptotic properties are systematically investigated. In longitudinal study, the classical maximum likelihood (ML) method has been extensively applied to analyze longitudinal missing data. To avoid sensitivity of ordinary ML estimates to extreme observations or outliers, Sinha (2012) suggested a robust method in the framework of the maximum likelihood for analyzing incomplete longitudinal data with generalized linear mixed models. Beyond that, Imai (2009) introduced an identification strategy for average treatment effect under the nonignorable assumption to analyze randomized experiments with a nonignorable missing binary outcome. In a sensitivity analysis, Xie et al. (2011) relaxed the linearity assumption for response probability and provided a semiparametric approach of the generalized additive model for analyzing nonignorable missing data. Their approach can avoid fitting any complicated semiparametric joint selection model. Lee and Tang (2006) considered a nonlinear structural equation model with nonignorable missing covariates and ordered categorical data, where the missingness mechanism was specified a logistic regression model.

As to nonignorable missing data, the underlying assumptions are difficult to verify in practice and the results of relevant statistical inference may be sensitive to these assumptions. Under this circumstance, parameter estimation for nonignorable nonresponse data is a challenge. To the best of our knowledge, few references focus on the inference for parameter β in linear regression with nonignorable missing response. The present paper focuses on this issue.

The empirical likelihood approach for constructing confidence intervals in nonparametric setting was introduced by Owen (1988, 1990). Since then, there has been a rich body of literature about relevant statistical inference based on the empirical likelihood technique. The empirical likelihood method owns its broad usage and widely research to a number of important advantages. As mentioned in Hall and La Scala (1990), the empirical likelihood technique does not impose prior constraints on the shape of the region and it does not require the construction of a pivotal quantity, besides, the region is range preserving and transformation respecting. Moreover, they are of natural shape and orientation since the regions are obtained by contouring a log-likelihood ratio. After that, Owen (1991) applied the empirical likelihood to linear regression and demonstrated that the empirical log-likelihood ratio is asymptotically a χ^2 variable. As to the construction of the confidence interval, Zhu and Xue (2006) studied the empirical likelihood-based inference for the parameters in a partially linear single-index model and first presented a bias correction to eliminate non-negligible bias caused by nonparametric estimation so as to achieve the standard χ^2-limit of the empirical likelihood function. For the missing data, Xue (2009a) developed an empirical likelihood method to study the construction of confidence intervals and regions for the parameters of interest in linear regression models with missing response data. Besides, Xue (2009b) elaborated the construction of the confidence interval for response mean based on the bias-corrected empirical likelihood ratio, where the missing response was imputed by a kernel regression method. Qin et al. (2009) raised a unified empirical likelihood approach for the case with the number of estimating equations greater than the number of unknown parameters.

The rest of this article is organized as follows. In Section 2, we present the construction of confidence intervals. The asymptotic normality for the estimates of the parameters and the asymptotic properties for the proposed empirical likelihood functions are investigated in Section 3. Simulation studies and a real data analysis are conducted to evaluate the finite sample performance of the proposed estimates in Sections 4 and 5, respectively. The concluding discussions are included in Section 6. Proofs of the asymptotic results are relegated in the Appendix.

2. Empirical likelihood-based inference

In this section, we propose three methods for the confidence interval construction of parameters in the following.

2.1. Weighted empirical likelihood

For an incomplete dataset $\{(x_i, y_i, \delta_i),\ i = 1, \dots, n\}$ with δ_i being the missing datum indicator with the response probability $p(x_i, y_i)$:

$$p(x_i, y_i) = P(\delta_i = 1 | x_i, y_i).$$

To construct the empirical likelihood function, the following auxiliary random vector based on the inverse probability weighted method is introduced:

$$z_{i,W} := z_{i,W}(\beta) = \frac{\delta_i}{p(x_i, y_i)} x_i (y_i - x_i^\tau \beta). \tag{2.1}$$

It is easy to verify that $E(z_{i,W}) = 0$. Let $(p_1, p_2, \ldots, p_n)^\tau$ be the vector of probability values, that is, $\sum_{i=1}^n p_i = 1$ and $p_i \geq 0$. Therefore, a weighted empirical likelihood function is defined as

$$L_W(\beta) = \sup\left\{\prod_{i=1}^n p_i | p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i z_{i,W} = 0\right\}.$$

Without the constraint $\sum_{i=1}^n p_i z_{i,W} = 0$, through the maximum of $\prod_{i=1}^n p_i$, we can obtain $p_i = 1/n$. The corresponding empirical likelihood ratio function for the parameter β is then given as

$$l_W(\beta) = -2\sup\left\{\sum_{i=1}^n \log(np_i) | p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i z_{i,W} = 0\right\}.$$

Further, the Lagrange multiplier method leads to

$$p_{i,W} = \frac{1}{n(1 + \lambda_W^\tau z_{i,W})}$$

where $\lambda_W \in \mathbb{R}^d$ is the Lagrange multiplier. The empirical log-likelihood ratio function for β can be shown to be

$$l_W(\beta) = 2\sum_{i=1}^n \log(1 + \lambda_W^\tau z_{i,W}). \tag{2.2}$$

Thus, the maximum empirical likelihood estimate (MELE) of $\hat{\beta}_W$ can be obtained by maximizing $-l_W(\beta)$ over all β. Under certain smooth conditions, $\hat{\beta}_W$ is the solution by simultaneously satisfying the following two equations: $T_{1n}(\hat{\beta}_W) = n^{-1} \sum_{i=1}^n z_{i,W}/(1+\lambda_W^\tau z_{i,W}) = 0$, and $T_{2n}(\hat{\beta}_W) = (2n)^{-1}\partial\{-l_W(\beta)\}/\partial\beta|_{\beta=\hat{\beta}_W} = n^{-1}\sum_{i=1}^n \lambda_W \delta_i x_i x_i^\tau/\{p(x_i, y_i)(1+\lambda_W^\tau z_{i,W})\} = 0$.

Note that the response probability $p(x_i, y_i)$ in (2.1) is regarded as a given value. When it is unknown with nonignorable missing data (NMAR), we refer to Kim and Yu (2011) to settle this issue. Suppose that the response probability model is a semiparametric logistic regression model

$$p(x_i, y_i) = \Pr(\delta_i = 1|x_i, y_i) = \frac{\exp\{g(x_i) + \phi y_i\}}{1 + \exp\{g(x_i) + \phi y_i\}}, \tag{2.3}$$

where $g(\cdot)$ is an unknown smooth function and ϕ is an unknown parameter. Thus, the conditional odds of nonresponse can be written as

$$O(x_i, y_i) = \frac{\Pr(\delta_i = 0|x_i, y_i)}{\Pr(\delta_i = 1|x_i, y_i)} = \frac{1}{p(x_i, y_i)} - 1. \tag{2.4}$$

Denote $\gamma = -\phi$ and together with model (2.3), we can further obtain $O(x_i, y_i) = \exp\{-g(x_i)+\gamma y_i\}$. Let $\alpha(x_i; \gamma) = O(x_i, y_i)/\exp(\gamma y_i)$. Under model (2.3), we have

$$\alpha(X; \gamma)E\{\delta \exp(\gamma Y)|X\} = E(1 - \delta|X) = E\{\delta O(X, Y)|X\}. \tag{2.5}$$

Based on the above equation, an estimate of $\alpha(x_i; \gamma)$ can be presented as

$$\hat{\alpha}(x_i; \gamma) = \frac{\sum_{j=1}^n (1 - \delta_j)\mathcal{K}_h(x_i, x_j)}{\sum_{j=1}^n \delta_j \exp(\gamma y_j)\mathcal{K}_h(x_i, x_j)} \tag{2.6}$$

where $\mathcal{K}_h(u, x) = h^{-1}\mathcal{K}\{(u-x)/h\}$ in which $\mathcal{K}(\cdot)$ is a symmetric kernel function and h is the bandwidth. Invoking Formula (2.4), we can obtain

$$\hat{p}(x_i, y_i) = \hat{p}(x_i, y_i; \gamma) = \{1 + \hat{\alpha}(x_i; \gamma)\exp(\gamma y_i)\}^{-1}. \tag{2.7}$$

It is worthwhile to mention that the parameter γ is not always known. As described in Rotnitzky et al. (1998), the parameter γ is assumed to be known in some cases, such as a sensitivity analysis or the planned missingness. However, in most of other scenarios, it is unknown and needs to be estimated. A parameter estimate $\hat{\gamma}$ can be obtained from an independent survey or a validation sample that is a subsample of the nonrespondents, whose estimate equation will be given in the following section.

2.2. Imputed empirical likelihood

For the above weighted empirical likelihood, the information contained in the data is not fully explored. Clearly, this may lead to a coverage accuracy deterioration of the confidence regions, especially in the case where missing data consist of a relatively large portion of all data. In order to make use of the known information as much as we can, two imputed methods are suggested to construct "completed datasets" as follows:

$$y_{i1}^{\star} = \delta_i y_i + (1-\delta_i)\hat{m}_0(x_i)$$
$$y_{i2}^{\star} = \frac{\delta_i}{\hat{p}(x_i, y_i)} y_i + \left\{1 - \frac{\delta_i}{\hat{p}(x_i, y_i)}\right\}\hat{m}_0(x_i)$$

where $m_0(x_i) = E(Y|x_i, \delta = 0)$ and $\hat{p}(x_i, y_i)$ is the estimate of $p(x_i, y_i)$, which has been defined in (2.7). For the semiparametric logistic regression model (2.3), we refer to Kim and Yu (2011) once again to get the estimate of $m_0(x_i)$ as follows:

$$\hat{m}_0(x_i) := \hat{m}_0(x_i; \gamma) = \sum_{j=1}^{n} \omega_{i0}(x_i; \gamma) y_j, \tag{2.8}$$

where the weight

$$\omega_{i0}(x_i; \gamma) = \frac{\delta_j \exp(\gamma y_j)\mathcal{K}_h(x_i, x_j)}{\sum_{j=1}^{n} \delta_j \exp(\gamma y_j)\mathcal{K}_h(x_i, x_j)}. \tag{2.9}$$

Based on the two complete datasets $\{(x_i, y_{i1}^{\star})\}$ and $\{(x_i, y_{i2}^{\star})\}$, we can construct the following auxiliary random vectors:

$$z_{i,k} := z_{i,k}(\beta) = x_i(y_{ik}^{\star} - x_i^{\tau}\beta), \quad k = 1, 2. \tag{2.10}$$

Consequently, the corresponding empirical log-likelihood ratio functions based on the imputed values can be defined as

$$l_{I,k}(\beta) = -2\sup\left\{\sum_{i=1}^{n} \log(np_i) | p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i z_{i,k} = 0\right\}, \quad k = 1, 2.$$

After that, through the Lagrange multiplier method, we can further obtain that

$$l_{I,k}(\beta) = 2\sum_{i=1}^{n} \log(1 + \lambda_{I,k}^{\tau} z_{i,k}), \quad k = 1, 2.$$

Therefore, two estimates $\hat{\beta}_{I,k},\ k = 1, 2$ can be acquired by maximizing $-l_{I,k}(\beta),\ k = 1, 2$.

Remark 1. In this paper, we focus on linear regression models with homoscedastic errors. It is worthwhile to note that the proposed methods are also readily extended to handle linear models with heteroscedastic errors. As an explanation, consider the following linear regression: $y_i = x_i^{\tau}\beta + \omega(x_i)\varepsilon_i,\ i = 1, \ldots, n$, where ε_i's are independent random errors with conditional mean $E(\varepsilon|X) = 0$. Because $E(\varepsilon|X) = 0$, we can obtain that the expectations $z_{i,W}$ and $z_{i,k},\ k = 1, 2$ in (2.1) and (2.10) respectively are still equal to zero. Thus our proposed methods can still yield consistent estimators and confidence regions. To improve the efficiencies, we need a proper weighting strategy. To be precise, when the form of $\omega(\cdot)$ is known, new auxiliary random vectors can be constructed by multiplying $z_{i,W}$ and $z_{i,k},\ k = 1, 2$ with $\omega^{-2}(x_i)$. When the function $\omega(\cdot)$ is also unknown, some nonparametric approach should be adopted. In the case with responses missing at random, Qin and Lei (2010) and Tang and Zhao (2013) studied the empirical likelihood inference in linear and nonlinear regression models with known $\omega(\cdot)$. We leave the theoretical study on this topic in a next research and instead we investigate the effect of heteroscedasticity on the proposed approaches in the simulation studies.

3. Theoretical results

In this section, for the cases with known and unknown γ, we provide the asymptotic results of the three estimates $\hat{\beta}_W,\ \hat{\beta}_{I,k},\ k = 1, 2$. From the proofs presented in the Appendix, we conclude that the three estimates all satisfy the asymptotic normality and the corresponding empirical likelihood functions are all asymptotically chi-squared.

3.1. Asymptotic properties of MELE with known $\gamma^{\star}$

We first consider the case with known $\gamma^{\star}$ that is the true value of parameter γ. In the following, denote $S = (X, Y)$ and $s_i = (x_i, y_i)$. For notational convenience, define

$$A = E\left[p^{-1}(S)XX^{\tau}\varepsilon^2 + \{1 - p^{-1}(S)\}XX^{\tau}E^2(\varepsilon|X, \delta = 0)\right],$$
$$T = E(XX^{\tau}), \qquad B_1 = E\left\{p^{-1}(S)XX^{\tau}\varepsilon^2\right\},$$
$$B_2 = E\left[p(S)XX^{\tau}\varepsilon^2 + \{1 - p(S)\}XX^{\tau}E^2(\varepsilon|X, \delta = 0)\right].$$

Theorem 1. *Suppose that Conditions* (C1)–(C6) *in the Appendix hold. Let β be the true value of the parameter. Then we have*

$$\text{(i) } \sqrt{n}(\hat\beta_W - \beta) \xrightarrow{D} N(0, T^{-1}AT^{-1}), \qquad \sqrt{n}(\hat\beta_{I,1} - \beta) \xrightarrow{D} N(0, T^{-1}AT^{-1}),$$
$$\sqrt{n}(\hat\beta_{I,2} - \beta) \xrightarrow{D} N(0, T^{-1}AT^{-1}),$$
$$\text{(ii) } \hat l_W(\beta, \gamma^\star) \xrightarrow{D} \rho_{1W}\chi^2_{1,1} + \rho_{2W}\chi^2_{1,2} + \cdots + \rho_{dW}\chi^2_{1,d},$$
$$\hat l_{I,1}(\beta, \gamma^\star) \xrightarrow{D} \rho_{1I}\chi^2_{1,1} + \rho_{2I}\chi^2_{1,2} + \cdots + \rho_{dI}\chi^2_{1,d}, \qquad \hat l_{I,2}(\beta, \gamma^\star) \xrightarrow{D} \chi^2_d$$

where $\chi^2_{1,i}$, $i = 1, \ldots, d$ s are independent and follow the standard χ^2 distribution with one degree of freedom, the weights ρ_{iW} and ρ_{iI} are the eigenvalues of $B_1^{-1}A$ and $B_2^{-1}A$ respectively, and χ^2_m means the χ^2 distribution with m degrees of freedom.

Theorem 1(i) characterizes the asymptotic normality of the three estimates $\hat\beta_W$ and $\hat\beta_{I,k}$, $k = 1, 2$, which can be applied to construct a normal-approximation-based (NA-based) confidence region for β at $1 - \alpha$ significance level. To be precise, denote $V_{\hat\beta_W} = T^{-1}AT^{-1}$, a NA-based confidence region of β with $1-\alpha$ level can be constructed by using the fact that $(\hat\beta_W - \beta)^\tau n\hat V^{-1}_{\hat\beta_W}(\hat\beta_W - \beta) \xrightarrow{D} \chi^2_d$ with $\hat V_{\hat\beta_W}$ being a consistent estimate of $V_{\hat\beta_W}$ which will be specified later. Further, Theorem 1(ii) presents the asymptotic distribution of the empirical likelihood functions $\hat l_W(\beta, \gamma^\star)$ and $l_{I,k}(\beta, \gamma^\star)$, $k = 1, 2$, which can also be used to test the hypothesis $H_0 : \beta = \beta_0$ and construct the confidence region for β. For example, a confidence region at the $1-\alpha$ level can be given by $CI_W(\tilde\beta) = \{\tilde\beta | \hat l_W(\tilde\beta, \gamma^\star) \le \sum_{i=1}^d \rho_{iW}\chi^2_1(1-\alpha)\}$, where $\chi^2_m(1-\alpha)$ is the $1-\alpha$ quantile of χ^2 distribution with the freedom of m degrees. One can reject the null hypothesis H_0 whenever $\hat l_W(\beta_0, \gamma^\star) > \sum_{i=1}^d \rho_{iW}\chi^2_1(1-\alpha)$.

To estimate the variances of $\hat\beta_W$ and $\hat\beta_{I,k}$, $k = 1, 2$, we need to obtain the estimates of A, T, B_1, and B_2. Based on the proof of Lemma 1 in the Appendix, the consistent estimates can be given by $\hat A = \sum_{i=1}^n \hat\theta_{A,i}\hat\theta^\tau_{A,i}/n$, $\hat B_1 = \sum_{i=1}^n \hat\theta_{B_1,i}\hat\theta^\tau_{B_1,i}/n$ and $\hat B_2 = \sum_{i=1}^n \hat\theta_{B_2,i}\hat\theta^\tau_{B_2,i}/n$, where

$$\hat\theta_{B_1,i} = \frac{\delta_i}{\hat p(s_i)} x_i(y_i - x_i^\tau\hat\beta),$$
$$\hat\theta_{B_2,i} = \delta_i x_i(y_i - x_i^\tau\hat\beta) + (1-\delta_i)x_i\hat E(\varepsilon|x_i, \delta = 0),$$
$$\hat\theta_{A,i} = \hat\theta_{B_1,i} + \left\{1 - \frac{\delta_i}{\hat p(s_i)}\right\} x_i\hat E(\varepsilon|x_i, \delta = 0),$$

here, $\hat E(\varepsilon|x_i, \delta = 0) = \sum_{j=1}^n \omega_{i0}(x_i, \gamma)(y_j - x_j\hat\beta)$ and $\hat\beta$ can be replaced by $\hat\beta_W$, $\beta_{I,k}$, $k = 1, 2$, respectively. At last, a consistent estimate of T can be defined to be $\hat T = \sum_{i=1}^n x_i x_i^\tau/n$.

3.2. Asymptotic properties of MELE with estimated $\hat\gamma$

When the true value $\gamma^\star$ is unknown, an estimate, say $\hat\gamma$, is required. In this article, the estimate $\hat\gamma$ comes from an independent survey or a validation sample. Here a validation sample is randomly selected from the set of non-respondents of the original sample. When a validation subsample is used to estimate the tilting parameter, we assume that the elements in the validation subsample are completely observed. Consider for example, in some survey about income, at the first stage, we get some missing responses due to various reasons. However, in a follow-up study, we get in touch with some nonrespondents in the first stage and then obtain their information. Hence, we can also call validation sample the follow-up sample. We use these two notions mutually. See more detailed information on validation subsample in Kim and Yu (2011) and Zhao et al. (2013). We first consider the case with an estimate $\hat\gamma$ from an independent survey with the sample size n. To state the theorem clearly, some notations are introduced as follows:

$$H = E\Big[X\{1 - p(S;\gamma)\}\{\varepsilon - E(\varepsilon|X, \delta = 0)\}^2\Big],$$
$$V = A + HH^\tau V_\gamma,$$

here V_γ is the variance of $\hat\gamma$.

Theorem 2. *Under Conditions* C1–C6 *in the Appendix, $\hat\gamma$ satisfies $\sqrt{n}(\hat\gamma - \gamma) \xrightarrow{D} N(0, V_\gamma)$ independent of both $n^{-1}\sum_{i=1}^n \hat z_{i,W}(\gamma)$ and $n^{-1}\sum_{i=1}^n z_{i,k}$, $k = 1, 2$. Let β be the true value of the parameter. Then we have*

$$\text{(i) } \sqrt{n}(\hat\beta_W - \beta) \xrightarrow{D} N(0, T^{-1}VT^{-1}), \qquad \sqrt{n}(\hat\beta_{I,1} - \beta) \xrightarrow{D} N(0, T^{-1}VT^{-1}),$$
$$\sqrt{n}(\hat\beta_{I,2} - \beta) \xrightarrow{D} N(0, T^{-1}VT^{-1}),$$
$$\text{(ii) } \hat l_W(\beta, \hat\gamma) \xrightarrow{D} \tilde\rho_{1W}\chi^2_{1,1} + \tilde\rho_{2W}\chi^2_{1,2} + \cdots + \tilde\rho_{dW}\chi^2_{1,d},$$

$$\hat{l}_{I,1}(\beta,\hat{\gamma}) \xrightarrow{D} \rho_{1,I1}\chi^2_{1,1} + \rho_{2,I1}\chi^2_{1,2} + \cdots + \rho_{d,I1}\chi^2_{1,d},$$
$$\hat{l}_{I,2}(\beta,\hat{\gamma}) \xrightarrow{D} \rho_{1,I2}\chi^2_{1,1} + \rho_{2,I2}\chi^2_{1,2} + \cdots + \rho_{d,I2}\chi^2_{1,d}$$

where $\chi^2_{1,i}$, $i = 1, \ldots, d$ s are independent and follow the standard χ^2 distribution with one degree of freedom, the weights $\tilde{\rho}_{iW}$, $\rho_{i,I1}$ and $\rho_{i,I2}$ are respectively the eigenvalues of $B_1^{-1}V$, $B_2^{-1}V$ and $A^{-1}V$.

Due to the estimation of γ, we have an extra dispersion term V_γ. A consistent estimate of V can be written as

$$\hat{V} = \hat{A} + \hat{H}\hat{H}^\tau \hat{V}_\gamma,$$

where $\hat{V}_\gamma = n\mathrm{Var}(\hat{\gamma})$ and the consistent estimate of H is given as

$$\hat{H} = \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)x_i\hat{\sigma}_0^2(x_i)$$

with

$$\hat{\sigma}_0^2(x_i) = \frac{\sum_{j=1}^{n}\delta_j \exp(\gamma y_j)\mathcal{K}_h(x_i,x_j)\{y_j - \hat{m}_0(x_j)\}^2}{\sum_{l=1}^{n}\delta_l \exp(\gamma y_l)\mathcal{K}_h(x_i,x_l)}.$$

Now we are in the position to consider an estimate from a validation sample, which is randomly selected from the set of nonrespondents and the responses are obtained for all the elements in the validation sample. We get the estimate $\hat{\gamma}$ for $\gamma^\star$ by solving the following equation:

$$\sum_{i=1}^{n}(1-\delta_i)r_i\{y_i - \hat{m}_0(x_i;\gamma)\} = 0 \tag{3.1}$$

where r_i is an indicator function, which takes 1 if unit i, $i = 1, \ldots, n$ belongs to the follow-up sample and otherwise, equals 0 and the expression of $\hat{m}_0(x_i;\gamma)$ has been defined in (2.8). In order to introduce the following theorem, we first give some notations as follows:

$$M = E\Big[r(1-\delta)\Big\{E(Y^2|x_i,\delta=0) - m_0^2(x_i;\gamma^\star)\Big\}\Big],$$
$$\eta_1 = \frac{\delta}{p(S)}X\varepsilon + \Big\{1 - \frac{\delta}{p(S)}\Big\}XE(\varepsilon|X,\delta=0) + HM^{-1}\Big[(1-\delta)r - \delta\nu\{p^{-1}(S)-1\}\Big]\{Y - m_0(X,\gamma^\star)\},$$

here $\nu = E(r|\delta = 0)$.

The following theorem states the asymptotic properties of the estimates $\hat{\beta}_W$ and $\hat{\beta}_{I,k}$, $k = 1, 2$. A sketch of the proof can be found in the Appendix.

Theorem 3. *Suppose that Conditions* C1–C6 *listed in the Appendix hold and assume that $\hat{\gamma}$ obtained from Eq. (3.1) exists almost everywhere. When β is the true value of the parameter, we have*

(i) $\sqrt{n}(\hat{\beta}_W - \beta) \xrightarrow{D} N(0, T^{-1}\widetilde{V}T^{-1})$, $\sqrt{n}(\hat{\beta}_{I,1} - \beta) \xrightarrow{D} N(0, T^{-1}\widetilde{V}T^{-1})$,
$\sqrt{n}(\hat{\beta}_{I,2} - \beta) \xrightarrow{D} N(0, T^{-1}\widetilde{V}T^{-1})$,

(ii) $\hat{l}_W(\beta,\hat{\gamma}) \xrightarrow{D} \check{\rho}_{1W}\chi^2_{1,1} + \check{\rho}_{2W}\chi^2_{1,2} + \cdots + \check{\rho}_{dW}\chi^2_{1,d}$,
$\hat{l}_{I,1}(\beta,\hat{\gamma}) \xrightarrow{D} \check{\rho}_{1,I1}\chi^2_{1,1} + \check{\rho}_{2,I1}\chi^2_{1,2} + \cdots + \check{\rho}_{d,I1}\chi^2_{1,d}$,
$\hat{l}_{I,2}(\beta,\hat{\gamma}) \xrightarrow{D} \check{\rho}_{1,I2}\chi^2_{1,1} + \check{\rho}_{2,I2}\chi^2_{1,2} + \cdots + \check{\rho}_{d,I2}\chi^2_{1,d}$

where $\chi^2_{1,i}$, $i = 1, \ldots, d$ s are independent and follow the standard χ^2 distribution with one degree of freedom, the weights $\check{\rho}_{iW}$, $\check{\rho}_{i,I1}$ and $\check{\rho}_{i,I2}$ are respectively the eigenvalues of $B_1^{-1}\widetilde{V}$, $B_2^{-1}\widetilde{V}$ and $A^{-1}\widetilde{V}$ and $\widetilde{V} = \mathrm{Var}(\eta_1)$.

To obtain the estimate $\hat{\widetilde{V}}$ for $\widetilde{V}$, the estimate $\hat{M}$ for M is given first:

$$\hat{M} = \frac{1}{n}\sum_{i=1}^{n} r_i(1-\delta_i)\Big\{\hat{E}(Y^2|x_i,\delta=0) - \hat{m}_0^2(x_i;\hat{\gamma})\Big\},$$

where $\hat{E}(Y^2|x_i,\delta=0) = \sum_{j=1}^{n}\omega_{i0}(x_i;\gamma)y_j^2$, here $\omega_{i0}(x_i;\gamma)$ has been defined in (2.9). Then we can obtain $\hat{\widetilde{V}}$ through the following formula:

$$\hat{\widetilde{V}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\eta}_{1,i}\hat{\eta}_{1,i}^\tau - \Big(\frac{1}{n}\sum_{i=1}^{n}\hat{\eta}_{1,i}\Big)\Big(\frac{1}{n}\sum_{i=1}^{n}\hat{\eta}_{1,i}\Big)^\tau,$$

where

$$\hat{\eta}_{1,i} = \frac{\delta_i}{\hat{p}(s_i)} x_i \varepsilon_i + \left\{1 - \frac{\delta_i}{\hat{p}(s_i)}\right\} x_i \hat{E}(\varepsilon | x_i, \delta = 0) + \hat{H}\hat{M}^{-1}\Big[(1-\delta_i) r_i - \delta_i \nu \{p^{-1}(s_i) - 1\}\Big]\{y_i - \hat{m}_0(x_i, \hat{\gamma})\},$$

here, $\hat{E}(\varepsilon | x_i, \delta = 0) = \sum_{j=1}^{n} \omega_{i0}(x_i; \gamma)(y_j - x_j^{\tau}\hat{\beta})$. Similar to the proof for Lemma 2 in the Appendix, we can obtain the consistency of $\hat{\tilde{V}}$.

Remark 2. We now discuss the bandwidth selection. For simplicity of bandwidth selection and ease of exposition, after standardization, a common bandwidth without the data-driven algorithm is often used for all variables. Kim and Yu (2011) and Zhao et al. (2013) adopted this strategy to use some ad-hoc approach: choosing $n^{-1/5}$ or $n^{-1/3}$ as the bandwidth. Recently, Köhler et al. (2014) gave a review and comparison of existing bandwidth selection methods for kernel regression. Three main approaches were discussed: the corrected average squared error (ASE) based method, the cross-validation family and the plug-in group methods. A comparison was conducted in numerical study. Xue (2009a) and Xue and Xue (2011) adopted the cross-validation method to select bandwidth involved in the regression problems with missing response at random. We follow this line and propose a new bandwidth selection method for the nonignorable missing response problem. To be precise, we choose a bandwidth by minimizing $CV(h) = n^{-1}\sum_{i=1}^{n} \delta_i(\delta_i - \hat{p}_{-i}(x_i, y_i))^2$. Here $\hat{p}_{-i}(x_i, y_i)$ is a 'leave-one-out' version of $\hat{p}(x_i, y_i)$.

The cross-validation in our setting is different from those with missing response at random. From (2.7), we can only obtain the response probability for those units with available responses. That is the reason why we add an additional term δ_i. Another alternative method is to use minimization over the following criterion $CV^*(h) = n^{-1}\sum_{i=1}^{n}(1-\delta_i) r_i\{y_i - \hat{m}_{0,-i}(x_i; \gamma)\}^2$. Here $\hat{m}_{0,-i}(x_i; \gamma)$ is a 'leave-one-out' version of $\hat{m}_0(x_i; \gamma)$. To save space, the detailed discussion is not included in this paper.

4. Simulation study

Simulation studies are conducted to examine our theory and to compare the performance of the empirical likelihood and of the normal approximation. In the following, we tabulate the simulation results for both the coverage probabilities (CP) and the average widths (AW) of the confidence intervals/regions.

4.1. One-dimensional cases

In this subsection, we first examine the performance of our proposed methods under the assumed exponential tilting model in (2.3). Then a robustness study towards some misspecifications of the selected model is conducted and finally the effect of heteroscedastic errors is studied.

For model (1.1), the dataset (x_i, y_i) is generated from $y_i = x_i + \varepsilon_i$, where the covariate X is generated from the normal distribution with mean 1 and variance 0.5 and the random error $\varepsilon_i \sim N(0, 0.25)$. In this circumstance, the parameter β is one-dimensional and the true value of β equals 1. Assume that the variable $X_i's$ is completely observed and some of $Y_i's$ subject to missingness. The response indicator variable δ_i follows the Bernoulli distribution with the probability p_i. The following four missing mechanisms are investigated:

Case 1. $p(x_i, y_i) = 1/\Big[1 + \exp\Big\{-(0.5x_i + 0.4x_i^2 + 0.3y_i)\Big\}\Big]$,

Case 2. $p(x_i, y_i) = 1/\Big[1 + \exp\Big\{-(-0.15 + 0.3x_i + 0.6y_i)\Big\}\Big]$,

Case 3. $p(x_i, y_i) = 1/\Big(1 + \exp\Big[-\{0.8\sin(x_i) + 0.6y_i\}\Big]\Big)$,

Case 4. $p(x_i, y_i) = 1/\Big(1 + \exp\Big[-\{0.3\exp(x_i) + 0.1y_i\}\Big]\Big)$.

As to the four preceding selection probabilities, the responding non-missing proportions are roughly 75.78%, 67.12%, 75.56% and 72.32%. Besides, the follow-up rate used is 30%. The estimated parameter $\hat{\gamma}$ is computed by solving (3.1) using the Newton–Raphson method. The kernel function $\mathcal{K}(u)$ is taken to be Gaussian kernel: $\mathcal{K}(u) = (2\pi)^{-1/2}\exp(-u^2/2)$. The cross-validation method are applied to select the optimal bandwidths h_{opt} based on two criteria proposed in Remark 2.

Monte Carlo relative biases, variances and mean squared errors (MSE) of the three point estimates $\hat{\beta}_W$, $\hat{\beta}_{I1}$ and $\hat{\beta}_{I2}$ for the above Cases 1–4 are summarized in Table 1. In addition, Tables 2 and 3 respectively present the coverage probability (CP) and average width (AW) of confidence intervals for β which are based on the empirical likelihood (EL) and the normal approximation (NA): the weighted empirical likelihood EL($\hat{\beta}_W$), the empirical likelihood based on simple imputed values EL($\hat{\beta}_{I1}$), the empirical likelihood based on weighted imputed values EL($\hat{\beta}_{I2}$) and the corresponding normal approximation approaches NA($\hat{\beta}_W$), NA($\hat{\beta}_{I1}$) and NA($\hat{\beta}_{I2}$). The sample sizes $n = 60, 70, 80, 90, 100$ are considered in the simulations and the significance level is set to be $\alpha = 0.05$. Every simulation result is the average of 5000 replications. In Tables 2 and 3, $p^{[1]}, p^{[2]}, p^{[3]}, p^{[4]}$ are used to signify the four missing mechanisms Cases 1–4.

From Table 1, we can have the following observations. First, the Monte Carlo relative biases, variances and mean squared errors (MSE) of the three point estimates $\hat{\beta}_W$, $\hat{\beta}_{I1}$ and $\hat{\beta}_{I2}$ are all small showing that the estimates perform well. Next,

Table 1
Relative biases, variances (Var) and mean squared errors (MSE) of three point estimates for Cases 1–4 in the one-dimensional case.

	n	Relative bias			Var			MSE		
		$\hat{\beta}_W$	$\hat{\beta}_{I1}$	$\hat{\beta}_{I2}$	$\hat{\beta}_W$	$\hat{\beta}_{I1}$	$\hat{\beta}_{I2}$	$\hat{\beta}_W$	$\hat{\beta}_{I1}$	$\hat{\beta}_{I2}$
$p^{[1]}$	60	−0.0087	−0.0140	−0.0128	0.0059	0.0077	0.0077	0.0060	0.0079	0.0078
	70	−0.0056	−0.0079	−0.0069	0.0027	0.0029	0.0028	0.0028	0.0029	0.0029
	80	−0.0054	−0.0077	−0.0067	0.0028	0.0029	0.0028	0.0028	0.0029	0.0028
	90	−0.0040	−0.0065	−0.0055	0.0013	0.0018	0.0017	0.0014	0.0018	0.0018
	100	−0.0036	−0.0055	−0.0046	0.0006	0.0007	0.0007	0.0006	0.0007	0.0007
$p^{[2]}$	60	−0.0017	−0.0106	−0.0071	0.0018	0.0029	0.0027	0.0018	0.0030	0.0028
	70	−0.0017	−0.0090	−0.0057	0.0010	0.0015	0.0014	0.0010	0.0016	0.0014
	80	−0.0026	−0.0084	−0.0053	0.0008	0.0010	0.0009	0.0008	0.0011	0.0009
	90	−0.0017	−0.0070	−0.0039	0.0008	0.0009	0.0008	0.0008	0.0010	0.0009
	100	−0.0016	−0.0068	−0.0039	0.0007	0.0009	0.0008	0.0007	0.0009	0.0008
$p^{[3]}$	60	−0.0077	−0.0171	−0.0146	0.0073	0.0094	0.0094	0.0074	0.0097	0.0096
	70	−0.0043	−0.0107	−0.0083	0.0029	0.0034	0.0034	0.0029	0.0036	0.0034
	80	−0.0028	−0.0087	−0.0062	0.0014	0.0019	0.0018	0.0014	0.0019	0.0018
	90	−0.0022	−0.0071	−0.0049	0.0011	0.0012	0.0011	0.0011	0.0013	0.0012
	100	−0.0018	−0.0062	−0.0042	0.0006	0.0008	0.0007	0.0006	0.0008	0.0007
$p^{[4]}$	60	−0.0030	−0.0072	−0.0058	0.0011	0.0020	0.0019	0.0011	0.0020	0.0019
	70	−0.0031	−0.0064	−0.0052	0.0016	0.0025	0.0024	0.0016	0.0025	0.0024
	80	−0.0028	−0.0056	−0.0044	0.0012	0.0016	0.0016	0.0012	0.0017	0.0016
	90	−0.0025	−0.0047	−0.0034	0.0007	0.0008	0.0007	0.0007	0.0008	0.0008
	100	−0.0021	−0.0042	−0.0032	0.0006	0.0007	0.0006	0.0006	0.0007	0.0006

Table 2
The coverage probabilities (CP) of confidence intervals for β based on the empirical likelihood (EL) and the normal approximation (NA) methods at the significance level $\alpha = 0.05$ for Cases 1–4 in the one-dimensional case.

	n	EL($\hat{\beta}_W$)	EL($\hat{\beta}_{I1}$)	EL($\hat{\beta}_{I2}$)	NA($\hat{\beta}_W$)	NA($\hat{\beta}_{I1}$)	NA($\hat{\beta}_{I2}$)
$p^{[1]}$	60	0.9532	0.9436	0.9480	0.9436	0.9235	0.9349
	70	0.9500	0.9419	0.9459	0.9429	0.9238	0.9372
	80	0.9508	0.9409	0.9445	0.9425	0.9221	0.9329
	90	0.9552	0.9403	0.9456	0.9449	0.9293	0.9376
	100	0.9517	0.9380	0.9433	0.9417	0.9277	0.9357
$p^{[2]}$	60	0.9568	0.9406	0.9456	0.9469	0.9122	0.9316
	70	0.9500	0.9388	0.9450	0.9426	0.9156	0.9320
	80	0.9508	0.9444	0.9512	0.9484	0.9218	0.9426
	90	0.9552	0.9340	0.9420	0.9393	0.9133	0.9290
	100	0.9543	0.9367	0.9393	0.9383	0.9140	0.9280
$p^{[3]}$	60	0.9532	0.9385	0.9442	0.9476	0.9176	0.9301
	70	0.9472	0.9305	0.9382	0.9362	0.9098	0.9222
	80	0.9538	0.9449	0.9469	0.9453	0.9229	0.9383
	90	0.9522	0.9346	0.9443	0.9436	0.9213	0.9340
	100	0.9534	0.9352	0.9418	0.9382	0.9186	0.9284
$p^{[4]}$	60	0.9523	0.9436	0.9506	0.9413	0.9209	0.9366
	70	0.9509	0.9419	0.9446	0.9413	0.9199	0.9329
	80	0.9523	0.9396	0.9430	0.9353	0.9223	0.9306
	90	0.9504	0.9400	0.9463	0.9440	0.9253	0.9363
	100	0.9507	0.9443	0.9508	0.9487	0.9313	0.9433

the three estimates are comparable for any specific missing mechanism. Further, it is reasonable that the relative biases, variances and mean squared errors decrease as the sample size n increases.

From Tables 2 and 3, it can be seen that all the coverage probabilities of the proposed approaches are very close to 0.95 at the significance level $\alpha = 0.05$. Besides, compared with the normal approximation, the empirical likelihood methods have uniformly higher coverage probabilities. Beyond that, the average widths of confidence intervals based on empirical likelihood methods are often shorter when the sample size is small or moderate, and are slightly longer but a little difference with large size samples. In other words, the empirical likelihood approaches get higher coverage probability and shorter interval when the sample size is small. In addition, among the empirical likelihood methods, EL($\hat{\beta}_W$) and EL($\hat{\beta}_{I2}$) are comparable and perform better than EL($\hat{\beta}_{I1}$) that has the lowest coverage probability and the longest average width.

We now conduct a simulation study to examine robustness of the proposed approaches towards misspecifications of the selected model. In this simulation, the same parameters settings as the above are used except the missing mechanisms, which are specified by the following four cases:

Table 3
The average widths (AW) of confidence intervals for β based on the empirical likelihood (EL) and the normal approximation (NA) methods at the significance level $\alpha = 0.05$ for Cases 1–4 in the one-dimensional case.

	n	EL($\hat{\beta}_W$)	EL($\hat{\beta}_{I1}$)	EL($\hat{\beta}_{I2}$)	NA($\hat{\beta}_W$)	NA($\hat{\beta}_{I1}$)	NA($\hat{\beta}_{I2}$)
$p^{[1]}$	60	0.1845	0.1893	0.1797	0.8652	0.8654	0.8653
	70	0.1311	0.1332	0.1305	0.6723	0.6724	0.6724
	80	0.1284	0.1296	0.1267	0.3524	0.3525	0.3524
	90	0.1152	0.1176	0.1136	0.1554	0.1555	0.1554
	100	0.1040	0.1058	0.1035	0.0977	0.0978	0.0976
$p^{[2]}$	60	0.2407	0.2657	0.2206	0.4935	0.4938	0.4936
	70	0.1793	0.1899	0.1668	0.3161	0.3163	0.3162
	80	0.1411	0.1503	0.1344	0.1549	0.1551	0.1550
	90	0.1292	0.1365	0.1242	0.1241	0.1242	0.1241
	100	0.1136	0.1161	0.1113	0.1041	0.1043	0.1042
$p^{[3]}$	60	0.2112	0.2188	0.2030	0.5253	0.5256	0.5254
	70	0.1698	0.1732	0.1637	0.3563	0.3565	0.3564
	80	0.1435	0.1472	0.1419	0.1584	0.1585	0.1584
	90	0.1363	0.1371	0.1337	0.1457	0.1458	0.1458
	100	0.1154	0.1180	0.1142	0.1008	0.1009	0.1009
$p^{[4]}$	60	0.1817	0.1936	0.1719	0.3956	0.3958	0.3957
	70	0.1418	0.1469	0.1396	0.2142	0.2144	0.2143
	80	0.1282	0.1322	0.1267	0.1365	0.1366	0.1365
	90	0.1144	0.1168	0.1131	0.1050	0.1051	0.1050
	100	0.1046	0.1063	0.1035	0.0992	0.0993	0.0992

Table 4
Relative biases, variances (Var) and mean squared errors (MSE) of three point estimates for Cases 5–8 in the one-dimensional case.

	n	Relative bias			Var			MSE		
		$\hat{\beta}_W$	$\hat{\beta}_{I1}$	$\hat{\beta}_{I2}$	$\hat{\beta}_W$	$\hat{\beta}_{I1}$	$\hat{\beta}_{I2}$	$\hat{\beta}_W$	$\hat{\beta}_{I1}$	$\hat{\beta}_{I2}$
$p^{[5]}$	60	−0.0279	−0.0432	−0.0418	0.0262	0.0353	0.0354	0.0270	0.0372	0.0371
	70	−0.0052	−0.0098	−0.0082	0.0030	0.0035	0.0035	0.0031	0.0036	0.0035
	80	−0.0047	−0.0093	−0.0078	0.0029	0.0033	0.0033	0.0029	0.0033	0.0033
	100	−0.0032	−0.0063	−0.0049	0.0009	0.0009	0.0009	0.0009	0.0010	0.0009
	150	−0.0023	−0.0058	−0.0043	0.0006	0.0007	0.0007	0.0006	0.0007	0.0007
$p^{[6]}$	60	−0.0022	−0.0087	−0.0066	0.0027	0.0045	0.0044	0.0027	0.0046	0.0045
	70	−0.0016	−0.0057	−0.0037	0.0012	0.0014	0.0013	0.0012	0.0014	0.0013
	80	−0.0009	−0.0043	−0.0025	0.0008	0.0009	0.0008	0.0008	0.0009	0.0008
	100	−0.0007	−0.0033	−0.0017	0.0006	0.0007	0.0006	0.0006	0.0007	0.0006
	150	−0.0001	−0.0019	−0.0006	0.0004	0.0005	0.0004	0.0004	0.0005	0.0004
$p^{[7]}$	60	−0.0014	−0.0048	−0.0026	0.0004	0.0005	0.0004	0.0004	0.0005	0.0004
	70	−0.0006	−0.0059	−0.0031	0.0008	0.0009	0.0009	0.0008	0.0010	0.0009
	80	−0.0006	−0.0056	−0.0029	0.0007	0.0008	0.0008	0.0007	0.0009	0.0008
	100	−0.0004	−0.0050	−0.0023	0.0007	0.0007	0.0007	0.0007	0.0008	0.0007
	150	−0.0001	−0.0037	−0.0014	0.0006	0.0006	0.0006	0.0006	0.0007	0.0006
$p^{[8]}$	60	−0.0143	−0.0186	−0.0175	0.0100	0.0133	0.0133	0.0107	0.0137	0.0136
	70	−0.0079	−0.0109	−0.0099	0.0044	0.0056	0.0055	0.0048	0.0057	0.0056
	80	−0.0050	−0.0097	−0.0086	0.0031	0.0048	0.0048	0.0031	0.0049	0.0048
	100	−0.0041	−0.0070	−0.0062	0.0022	0.0030	0.0029	0.0022	0.0030	0.0030
	150	−0.0032	−0.0045	−0.0039	0.0007	0.0008	0.0007	0.0007	0.0008	0.0008

Case 5. $p(x_i, y_i) = 1/\Big[1 + \exp\Big\{-(-0.35 + x_i + 0.8y_i + 0.1y_i^2)\Big\}\Big]$,

Case 6. $p(x_i, y_i) = 1/\Big[1 + \exp\Big\{-(-0.15 + 0.1x_i + 0.6y_i + 0.9x_iy_i)\Big\}\Big]$,

Case 7. $p(x_i, y_i) = 1 - \exp\Big\{-\exp(-0.05 + 0.3y_i)\Big\}$,

Case 8. $p(x_i, y_i) = \Phi(0.6x_i + 0.1x_i^2 + 0.3y_i - 0.1)$.

Here, 5000 random Monte Carlo samples of sizes $n = 60, 70, 80, 100, 150$ are generated for each of the above selection probability functions. For the above four cases, the mean response rates are 78.53%, 73.06%, 72.07% and 79.66%, respectively. The simulation results are tabulated in Tables 4–6, where $p^{[5]}, p^{[6]}, p^{[7]}, p^{[8]}$ are specified to indicate Cases 5–8. From these three tables, we can see clearly that even though the missingness mechanism is misspecified under the above four different cases, the three point estimates $\hat{\beta}_W$, $\hat{\beta}_{I1}$ and $\hat{\beta}_{I2}$ are all close to the true value. Moreover, the empirical coverage probabilities of both the empirical likelihood approaches and the normal approximation methods are close to the pre-specified nominal level 95%, but the former performs better than the latter. As for average widths of confidence intervals, we can observe that

Table 5
The coverage probabilities (CP) of confidence intervals for β based on the empirical likelihood (EL) and the normal approximation (NA) methods at the significance level $\alpha = 0.05$ for Cases 5–8 in the one-dimensional case.

	n	EL($\hat{\beta}_W$)	EL($\hat{\beta}_{I1}$)	EL($\hat{\beta}_{I2}$)	NA($\hat{\beta}_W$)	NA($\hat{\beta}_{I1}$)	NA($\hat{\beta}_{I2}$)
$p^{[5]}$	60	0.9402	0.9416	0.9402	0.9357	0.9263	0.9284
	70	0.9499	0.9489	0.9499	0.9459	0.9228	0.9349
	80	0.9502	0.9442	0.9465	0.9469	0.9271	0.9358
	100	0.9536	0.9367	0.9453	0.9490	0.9257	0.9387
	150	0.9506	0.9364	0.9430	0.9426	0.9274	0.9360
$p^{[6]}$	60	0.9515	0.9482	0.9502	0.9458	0.9301	0.9361
	70	0.9557	0.9430	0.9466	0.9430	0.9236	0.9370
	80	0.9527	0.9417	0.9483	0.9417	0.9247	0.9383
	100	0.9517	0.9453	0.9503	0.9507	0.9333	0.9427
	150	0.9503	0.9400	0.9470	0.9437	0.9323	0.9410
$p^{[7]}$	60	0.9472	0.9291	0.9352	0.9322	0.9074	0.9228
	70	0.9613	0.9437	0.9473	0.9440	0.9237	0.9333
	80	0.9520	0.9497	0.9543	0.9487	0.9277	0.9423
	100	0.9517	0.9387	0.9437	0.9390	0.9233	0.9340
	150	0.9520	0.9433	0.9517	0.9517	0.9330	0.9457
$p^{[8]}$	60	0.9544	0.9519	0.9532	0.9506	0.9266	0.9380
	70	0.9471	0.9390	0.9424	0.9404	0.9203	0.9310
	80	0.9495	0.9431	0.9444	0.9398	0.9267	0.9354
	100	0.9546	0.9445	0.9482	0.9455	0.9338	0.9412
	150	0.9440	0.9340	0.9420	0.9373	0.9273	0.9350

Table 6
The average widths (AW) of confidence intervals for β based on the empirical likelihood (EL) and the normal approximation (NA) methods at the significance level $\alpha = 0.05$ for Cases 5–8 in the one-dimensional case.

	n	EL($\hat{\beta}_W$)	EL($\hat{\beta}_{I1}$)	EL($\hat{\beta}_{I2}$)	NA($\hat{\beta}_W$)	NA($\hat{\beta}_{I1}$)	NA($\hat{\beta}_{I2}$)
$p^{[5]}$	60	0.2266	0.2311	0.2261	0.6865	0.6867	0.6867
	70	0.1836	0.1827	0.1799	0.4162	0.4163	0.4163
	80	0.1801	0.1813	0.1778	0.2501	0.2503	0.2503
	100	0.1258	0.1268	0.1253	0.1657	0.1658	0.1658
	150	0.0778	0.0779	0.0778	0.0811	0.0818	0.0809
$p^{[6]}$	60	0.1966	0.2056	0.1895	0.4456	0.4458	0.4457
	70	0.1517	0.1559	0.1509	0.4106	0.4108	0.4107
	80	0.1308	0.1336	0.1301	0.3460	0.3463	0.3461
	100	0.1556	0.1557	0.1556	0.1117	0.1135	0.1112
	150	0.0802	0.0803	0.0802	0.0827	0.0836	0.0824
$p^{[7]}$	60	0.6194	0.6190	0.5858	0.2236	0.2381	0.2003
	70	0.1619	0.1745	0.1548	0.3024	0.3027	0.3022
	80	0.1315	0.1414	0.1265	0.2223	0.2224	0.02223
	100	0.1114	0.1157	0.1093	0.1010	0.1011	0.1010
	150	0.0861	0.0868	0.0849	0.0824	0.0825	0.0824
$p^{[8]}$	60	0.2045	0.2130	0.2081	0.2682	0.2683	0.2683
	70	0.1789	0.1816	0.1795	0.2239	0.2241	0.2240
	80	0.1398	0.1452	0.1429	0.1679	0.1680	0.1681
	100	0.1110	0.1123	0.1107	0.1058	0.1059	0.1058
	150	0.0791	0.0797	0.0790	0.0769	0.0770	0.0769

when the sample sizes are small such as $n = 60, 70, 80$, the normal approximations suffer from long average widths of the confidence intervals, whereas the empirical likelihood methods can still control them reasonably. When the sample size gets relatively large, both the methods perform well and can obtain similar results. This study suggests the robustness of the proposed approaches and also give more information on the differences between the empirical likelihood and normal approximation. We can conclude that EL($\hat{\beta}_W$) and EL($\hat{\beta}_{I2}$) are comparable and more recommendable, which have higher coverage probabilities and shorter widths than EL($\hat{\beta}_{I1}$) has.

Further, since heteroscedastic errors are very common in practice, we also consider two heteroscedastic circumstances based on two missing mechanisms: Cases 1 and 5. Except for $\varepsilon_i \sim N(0, \sigma_i^2)$ in $y_i = x_i + \varepsilon_i$, where $\sigma_i^2 = 0.25x_i^2$, other parameters settings are the same as those in the homoscedastic situations. The corresponding simulation results are reported in Table 7. To save space, we only display coverage probabilities and average widths. From the results, we can see that the coverage probabilities of both the empirical likelihood methods and normal approximation approaches are close to the nominal level 95% and uniformly, for the coverage probabilities, the former works better than the latter. Another important observation here is that the average widths with heteroscedastic errors are longer than those in homoscedastic circumstances,

Table 7
Coverage probabilities (CP) and average widths (AW) for β with heteroscedastic errors based on the empirical likelihood (EL) and the normal approximation (NA) methods at the significance level $\alpha = 0.05$.

		n	EL($\hat{\beta}_W$)	EL($\hat{\beta}_{I1}$)	EL($\hat{\beta}_{I2}$)	NA($\hat{\beta}_W$)	NA($\hat{\beta}_{I1}$)	NA($\hat{\beta}_{I2}$)
$p^{[1]}$	CP	60	0.9514	0.9497	0.9484	0.9252	0.9172	0.9216
		70	0.9509	0.9520	0.9498	0.9325	0.9291	0.9311
		80	0.9530	0.9500	0.9496	0.9370	0.9316	0.9343
		100	0.9520	0.9477	0.9493	0.9323	0.9300	0.9323
		150	0.9467	0.9410	0.9427	0.9360	0.9330	0.9343
	AW	60	0.3805	0.3828	0.3730	0.7261	0.7266	0.7256
		70	0.3312	0.3323	0.3263	0.5665	0.5670	0.5660
		80	0.3037	0.3050	0.3007	0.5032	0.5033	0.5026
		100	0.2604	0.2612	0.2598	0.3285	0.3288	0.3219
		150	0.2033	0.2049	0.2016	0.1976	0.1965	0.1967
$p^{[5]}$	CP	60	0.9326	0.9360	0.9370	0.9249	0.9205	0.9242
		70	0.9457	0.9514	0.9521	0.9384	0.9337	0.9357
		80	0.9519	0.9502	0.9485	0.9382	0.9308	0.9325
		100	0.9520	0.9527	0.9513	0.9443	0.9380	0.9430
		150	0.9510	0.9503	0.9497	0.9413	0.9363	0.9383
	AW	60	0.5151	0.5167	0.5092	0.7195	0.7194	0.7189
		70	0.4366	0.4412	0.4340	0.6886	0.6883	0.6880
		80	0.3774	0.3715	0.3672	0.5337	0.5338	0.5334
		100	0.2956	0.2953	0.2929	0.3176	0.3179	0.3172
		150	0.2159	0.2158	0.2141	0.1999	0.1995	0.1998

especially for the cases with small and moderate sample sizes. It is reasonable that with the sample size increasing, the difference is decreasing. Moreover, for heteroscedastic cases, EL($\hat{\beta}_W$) and EL($\hat{\beta}_{I2}$) again have comparable performance and are more recommendable than EL($\hat{\beta}_{I1}$).

All in all, we can conclude that (i) the empirical likelihood performs well with small and moderate sample sizes, and works similarly to the normal approximation for the case with large sample size; (ii) EL($\hat{\beta}_W$) and EL($\hat{\beta}_{I2}$) are comparable and outperform EL($\hat{\beta}_{I1}$).

4.2. Two-dimensional cases

Consider the linear model (1.1) with $d = 2$ and the true value $\beta = (0.8, 1.0)$. In Model A, both x_{i1} and x_{i2} are independently generated from the uniform distribution $U(0, 1)$ and the error ε_i is generated from the normal distribution with mean zero and variance 0.1. The selection probability function is taken to be $p_i(x_{i1}, x_{i2}, y_i) = 1/[1+\exp\{-(1+0.25x_{i1}+0.25x_{i2}+0.15y_i)\}]$ and the average non-missing ratio is 79.86%. The kernel function $\mathcal{K}(u_1, u_2) = \mathcal{K}_0(u_1)\mathcal{K}_0(u_2)$ is the product kernel, where $\mathcal{K}_0(u) = (2\pi)^{-1/2}\exp(-u^2/2)$ and the optimal bandwidth h_{opt} is selected through the cross-validation. The sample size $n = 100$. In Model B, the settings are the same as Model A except that x_{i1} and x_{i2} are dependent. To be specific, x_{i1} comes from the uniform distribution $U(0, 1)$ and x_{i2} is generated from the model: $x_{i2} = x_{i1} + \epsilon_i$, where $\epsilon_i \sim U(0, 1) - 0.5$. Under the same selection probability function as Model A, the non-missing proportion is 78.94%. As for Model A and Model B, the confidence intervals for parameter β and their coverage probabilities are calculated from 2000 replications, which are based on the empirical likelihood methods EL($\hat{\beta}_W$), EL($\hat{\beta}_{I1}$), EL($\hat{\beta}_{I2}$) and the normal approximation approach NA($\hat{\beta}_W$), NA($\hat{\beta}_{I1}$), NA($\hat{\beta}_{I2}$). The simulation results about Model A and Model B are respectively reported in Figs. 1 and 2.

Fig. 1(a) is for Model A with the empirical likelihood method. From Fig. 1(a), the differences between the three empirical likelihood-based confidence regions are minor although, slightly but visibly, EL($\hat{\beta}_{I2}$) is the most efficient, followed by EL($\hat{\beta}_W$) and EL($\hat{\beta}_{I1}$) is the worst. Besides, the coverage probability of EL($\hat{\beta}_{I2}$) is 92.82%, whilst those for EL($\hat{\beta}_W$) and EL($\hat{\beta}_{I1}$) are 91.60% and 90.53%, respectively. The simulation results of Model A on the basis of the normal approximation are presented in Fig. 1(b), and the conclusions on the comparison between the three NA methods are almost the same as the above. That is, NA($\hat{\beta}_{I2}$) gives the smallest region, NA($\hat{\beta}_W$) comes the second and NA($\hat{\beta}_{I1}$) is the worst. In addition, the coverage probabilities of NA($\hat{\beta}_W$), NA($\hat{\beta}_{I1}$) and NA($\hat{\beta}_{I2}$) are 90.90%, 90.01% and 91.92%, respectively. Thus, NA($\hat{\beta}_{I2}$) is the most efficient. Comparing Fig. 1(a) with Fig. 1(b), the empirical likelihood has higher coverage probability correspondingly. We also make a comparison between the corresponding confidence regions, the empirical likelihood-based methods perform slightly worse. Thus overall, the two methodologies are comparable.

The simulation results of Model B based on the empirical likelihood and the normal approximation are respectively reported in Fig. 2(a) and Fig. 2(b). According to Fig. 2(a), all the three methods have almost the same regions. For coverage probabilities, EL($\hat{\beta}_{I2}$) gets 90.21% while those for EL($\hat{\beta}_{I1}$) and EL($\hat{\beta}_W$) are 89.30% and 87.62%. Thus, again, EL($\hat{\beta}_{I2}$) performs the best. For the normal approximation, the situation is different. Fig. 2(b) indicates that, very slightly, NA($\hat{\beta}_{I2}$) is the best, NA($\hat{\beta}_W$) the second and NA($\hat{\beta}_{I1}$) the worst. However, the corresponding coverage probabilities of NA($\hat{\beta}_W$), NA($\hat{\beta}_{I1}$) and

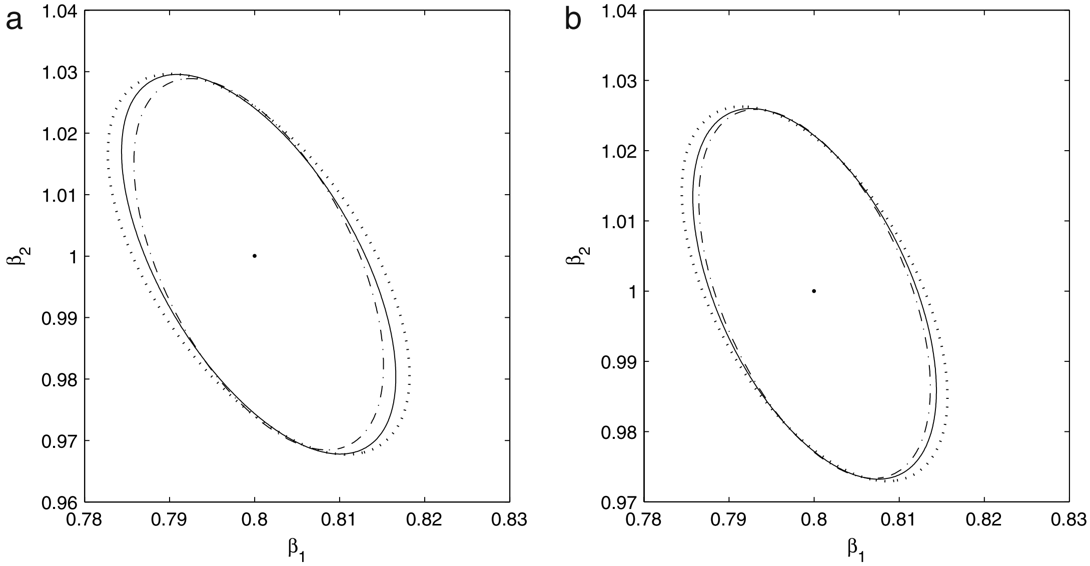

Fig. 1. The simulation results in the two-dimensional case. The 95% confidence regions for β: the left plot (a) and the right plot (b) are the results of Model *A* based on the empirical likelihood method and the normal approximation approach, respectively. In plot (a), the solid line is for EL($\hat{\beta}_W$), the dotted line for EL($\hat{\beta}_{I1}$) and the dash–dot line for EL($\hat{\beta}_{I2}$). In plot (b), the solid line, the dotted line and the dash–dot line are for NA($\hat{\beta}_W$), NA($\hat{\beta}_{I1}$) and NA($\hat{\beta}_{I2}$), respectively.

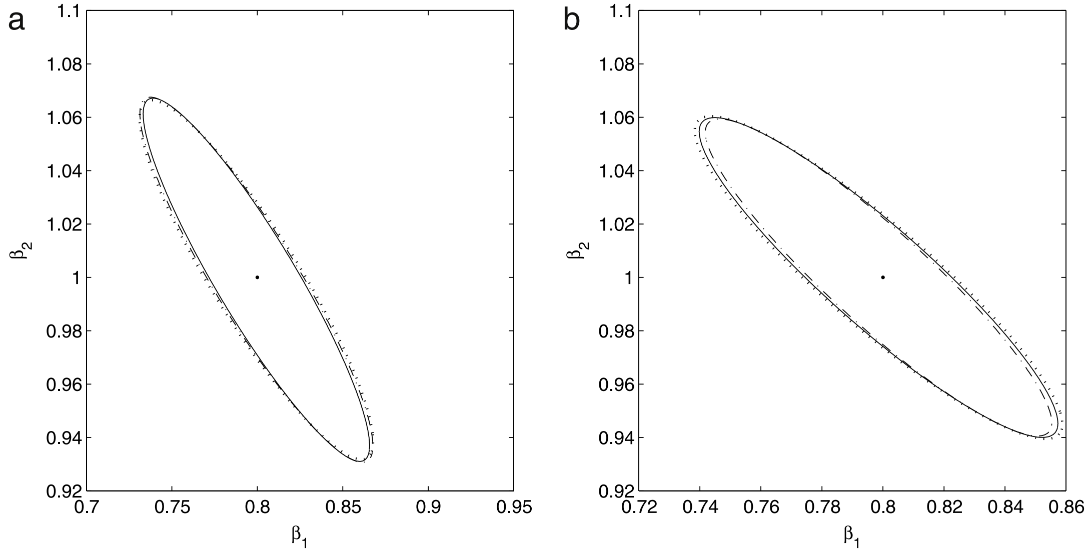

Fig. 2. The simulation results in the two-dimensional case. The 95% confidence regions for β: the left-hand plot (a) and the right-hand plot (b) are the results of Model *B* based on the empirical likelihood method and the normal approximation, respectively. In plot (a), the solid line is for EL($\hat{\beta}_W$), the dotted line for EL($\hat{\beta}_{I1}$) and the dash–dot line for EL($\hat{\beta}_{I2}$). In plot (b), the solid line, the dotted line and the dash–dot line are for NA($\hat{\beta}_W$), NA($\hat{\beta}_{I1}$) and NA($\hat{\beta}_{I2}$), respectively.

NA($\hat{\beta}_{I2}$) are 92.70%, 87.31%, 89.82%. Thus, NA($\hat{\beta}_W$) is the best and NA($\hat{\beta}_{I2}$) the worst. This suggests that the normal approximation has different performances according to different models, while the empirical likelihood is more stable. Comparing Fig. 2(a) with Fig. 2(b), we can conclude the normal approximation has slightly smaller region and lower coverage probabilities than the empirical likelihood methods have.

In summary, the empirical likelihood and the normal approximation have similar performance, and the former is more stable against model than the normal approximation is.

5. Real data analysis

We now apply the proposed methods to analyze a subsample of data on the persistence of maternal smoking from the Six Cities Study of the health effects of air pollution. The data consist of 574 observations where the outcome variable is

Table 8
The 95% confidence intervals of parameter β for the real data.

EL		NA	
EL($\hat{\beta}_W$)	(0.1198, 0.3536)	NA($\hat{\beta}_W$)	(0.1217, 0.3554)
EL($\hat{\beta}_{I1}$)	(0.1186, 0.3525)	NA($\hat{\beta}_{I1}$)	(0.1216, 0.3545)
EL($\hat{\beta}_{I2}$)	(0.1225, 0.3532)	NA($\hat{\beta}_{I2}$)	(0.1278, 0.3575)

a measure of the mother's smoking (in cigarettes per day) when her child is 10 years old, of these, the 208 subjects have missing outcome data. A more detailed description on this dataset can be seen in Ware et al. (1984). In addition, Lipsitz et al. (2004) applied this dataset to estimate the parameters in a linear regression, where the outcome is the square root of maternal smoking when the child is 10 years old (Smoke_{i2}) and the covariates are square root of maternal smoking when the child is 9 (Smoke_{i1}), the child's wheeze status at age 9 (Wheeze_i) and the city of residence (City_i). In their paper, the 208 missing responses are regarded as nonignorably missing Gaussian outcomes and the three covariates are all observed. However, their analysis result shows that the latter two covariates are not significant, thus, our interest is to fit the linear regression on the response Smoke_{i2} and the covariate Smoke_{i1}:

$$E(\text{Smoke}_{i2}|\text{Smoke}_{i1}) = \beta\text{Smoke}_{i1},$$

where

$$\text{Smoke}_{i2} = \sqrt{\text{maternal cigarettes smoked per day when the child is 10}},$$
$$\text{Smoke}_{i1} = \sqrt{\text{maternal cigarettes smoked per day when the child is 9}}.$$

An estimate $\hat{\gamma}$ of the exponential tilting parameter γ can be obtained by solving $\sum_{i=1}^{n}(1-\delta_i)r_i\{y_i - \hat{m}_0(x_i;\gamma)\} = 0$ in (3.1), where $\hat{m}_0(x_i;\gamma)$ is defined by (2.8) and r_i is an indicator variable which equals 1 when the ith unit belongs to the follow-up sample and 0 otherwise. The follow-up rate used here is 30%. To get $\hat{m}_0(x_i;\gamma)$, the Gaussian kernel function $\mathcal{K}(u) = (2\pi)^{-1/2}\exp(-u^2/2)$ is applied and the optimal bandwidth h_{opt} is chosen by the cross-validation method. Here, we employ the centralized data and the values of the three point estimates are $\hat{\beta}_W = 0.2385$, $\hat{\beta}_{I1} = 0.2385$, $\hat{\beta}_{I2} = 0.2406$. Besides, the 95% confidence intervals of β are reported in Table 8. As the conclusions we achieve from the above simulation section, when the sample size is large, the performance of both the empirical likelihood and normal approximation is comparable, which is verified by the results in Table 8.

6. Discussions

In this paper, the empirical likelihood and the exponential tilting model are used to construct confidence intervals for the parameters of interest in linear regression models with nonignorable missing response data. In line with the above simulation results, we can see the advantages and disadvantages of our proposed approaches as follows. In terms of coverage probabilities, the empirical likelihood methods uniformly outperform the corresponding normal approximation approaches. With regard to average widths, the empirical likelihood-based confidence intervals are shorter than the normal approximation-based ones for small or moderate sample size, and are slightly longer for large sample size. Overall, the empirical likelihood methods perform well and the normal approximation can also be recommendable when the sample size is large and its easy implementation is taken into consideration.

Note that with small and moderate sample sizes, the size of the validation sample could also be small. This could make the exponential tilting model difficult to obtain stable and accurate estimator for the tilting parameter γ. Another drawback of this semiparametric approach is that it can be affected by dimensionality significantly when the dimension of X is relatively high, see also Kim and Yu (2011) for this point. To deal with this problem, a possible approach to modify the exponential tilting model is as follows:

$$p(x_i, y_i) = \Pr(\delta_i = 1|x_i, y_i) = \frac{\exp\{g(\theta^\tau x_i) + \phi y_i\}}{1 + \exp\{g(\theta^\tau x_i) + \phi y_i\}}. \tag{6.1}$$

Compared with model (2.3), a dimension-reduction structure with an additional parameter θ is considered. If the true value of θ is given in advance, the problem is solved. When it is unknown, a Bayesian argument would be applied to have a prior about θ. Investigation on alternative methods for estimating θ, including Bayesian approach, deserves a further research. Of course, we may also consider a purely parametric model:

$$p(x_i, y_i) = \Pr(\delta_i = 1|x_i, y_i) = \frac{\exp\{\gamma_0 + \gamma_1^\tau x_i + \gamma_2 y_i\}}{1 + \exp\{\gamma_0 + \gamma_1^\tau x_i + \gamma_2 y_i\}}. \tag{6.2}$$

The estimation may be then relatively easier. The above observation says that when some more information about the response probability model is available, we can alleviate the dimensionality issue in certain sense.

On the other hand, when the response is nonignorable missing, it is not easy to check the identifiability of the parameter in parametric response mechanism. Moreover, the statistical inference could be sensitive to the failure of assumed parametric models. Accounting for this, semiparametric selection model would yield more robust results. Relevant theoretical analysis of the efficiency and robustness of the exponential tilting model over a purely parametric model would be an interesting and important topic. It is left to a future research topic.

As commonly used in the missing data analysis with nonignorable missing mechanism, see for example Jamshidian and Yuan (2013) and Noémie et al. (2011), sensitivity analysis is of importance, it deserves a further investigation though we very briefly discuss it in the present paper.

Acknowledgments

The research was supported by the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 14XNH102), National Natural Science Foundation of China (No. 11071253) and a grant from the Research Grants Council of Hong Kong (No. HKBU2040/13P). The authors thank the editor, the associate editor and two anonymous referees for their constructive comments and suggestions which led to a substantial improvement of an early manuscript.

Appendix. Proofs of theorems

The following conditions are required for proving the theorems in Section 3.

(C1) The probability density function $f(x)$ is bounded away from ∞ in the support of X and the first and second derivatives of $f(x)$ are continuous, smooth and bounded.
(C2) The kernel function $\mathcal{K}(\cdot)$ is a probability density function such that
  (a) it is bounded and has a compact support;
  (b) it is symmetric with $\sigma_k^2 = \int u^2 \mathcal{K}(u)du < \infty$;
  (c) $\mathcal{K}(u) \geq d_0$ for some $d_0 > 0$ in some closed interval centered at zero.
(C3) The bandwidth h satisfies: $nh \to \infty$ and $nh^4 \to \infty$ as $n \to \infty$.
(C4) The response probability function $p(x, y)$ satisfies: $p(x, y) > c_0 > 0$ for a positive constant c_0 and $p(X) = E\{p(X, Y)|X\} \neq 1$ almost surely.
(C5) $E(Y^2)$ and $E\{\exp(2\gamma Y)\}$ are finite.
(C6) The matrices A, B_1 and B_2 that are defined in Theorem 1 are positive definite.

Remark 3. Condition (C1) is required in probability theory. Conditions (C2)–(C3) are the common requisites for the kernel density estimation problem. Conditions (C4)–(C5) are similar to the conditions in Kim and Yu (2011). Condition (C6) is necessary for asymptotic normality.

The following lemmas are useful for proving theorems given in Section 3.

Lemma 1. *Under conditions* C1–C6 *in the Appendix, if β is the true parameter and when the parameter γ is known, we then have*

$$\text{(a)}\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,W}(\beta) \xrightarrow{D} N(0, A), \qquad \frac{1}{n}\sum_{i=1}^{n} z_{i,W}(\beta) z_{i,W}^{\tau}(\beta) \xrightarrow{P} B_1, \tag{A.1}$$

$$\text{(b)}\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,1}(\beta) \xrightarrow{D} N(0, A), \qquad \frac{1}{n}\sum_{i=1}^{n} z_{i,1}(\beta) z_{i,1}^{\tau}(\beta) \xrightarrow{P} B_2, \tag{A.2}$$

$$\text{(c)}\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,2}(\beta) \xrightarrow{D} N(0, A), \qquad \frac{1}{n}\sum_{i=1}^{n} z_{i,2}(\beta) z_{i,2}^{\tau}(\beta) \xrightarrow{P} A, \tag{A.3}$$

where the above notations have been defined in Section 3.1 in detail.

Proof of Lemma 1. (a) We first prove Lemma 1 for $z_{i,W}(\beta)$. Invoking Formula (2.1) and using the same notations $S = (X, Y)$ and $s_i = (x_i, y_i)$ in Section 3, we have

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,W}(\beta) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\delta_i}{p(s_i)} x_i\varepsilon_i + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\delta_i}{\hat{p}(s_i)} - \frac{\delta_i}{p(s_i)}\right\} x_i\varepsilon_i \\ &:= J_1 + J_2.\end{aligned}$$

Note that

$$\begin{aligned}\hat{p}(s_i) &:= \hat{p}(x_i, y_i) = \{1 + \hat{\alpha}(x_i; \gamma)\exp(\gamma y_i)\}^{-1}, \\ p(s_i) &:= p(x_i, y_i) = \{1 + \alpha(x_i; \gamma)\exp(\gamma y_i)\}^{-1},\end{aligned}$$

where $\hat{\alpha}(x_i;\gamma)$ has been defined in (2.6). From Kim and Yu (2011), we can get $n^{-1}\sum_{j=1}^{n}\delta_j\exp(\gamma y_j)\mathcal{K}_h(x_i,x_j) = f(x_i)\{1-p(x_i)\}\alpha^{-1}(x_i;\gamma)+o_p(1)$, where $f(\cdot)$ is the marginal density of X and $p(x) = E(\delta|x)$. In the following, we turn to consider the term J_2:

$$\begin{aligned}
J_2 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_i x_i\varepsilon_i\exp(\gamma y_i)\frac{\sum_{j=1}^{n}\{(1-\delta_j)-\delta_j\exp(\gamma y_j)\alpha(x_i;\gamma)\}\mathcal{K}_h(x_i,x_j)}{\sum_{j=1}^{n}\delta_j\exp(\gamma y_j)\mathcal{K}_h(x_i,x_j)}\\
&= \frac{1}{n^{3/2}}\sum_{i=1}^{n}\delta_i x_i\varepsilon_i\exp(\gamma y_i)\frac{\sum_{j=1}^{n}\{(1-\delta_j)-\delta_j\exp(\gamma y_j)\alpha(x_i;\gamma)\}\mathcal{K}_h(x_i,x_j)}{f(x_i)\{1-p(x_i)\}\alpha^{-1}(x_i;\gamma)}+o_p(1)\\
&= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\{(1-\delta_j)-\delta_j O(x_j,y_j)\}E\left[\frac{\delta\varepsilon X O(X,Y)\mathcal{K}_h(X,x_j)}{f(X)\{1-p(X)\}}\bigg|x_j\right]+o_p(1)\\
&= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}x_j\left\{1-\frac{\delta_j}{p(x_j,y_j)}\right\}\frac{E\{\varepsilon(1-\delta)|x_j\}}{E(1-\delta|x_j)}+o_p(1)\\
&= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}x_j\left\{1-\frac{\delta_j}{p(x_j,y_j)}\right\}E(\varepsilon|x_j,\delta=0)+o_p(1).
\end{aligned}$$

The penultimate equation follows from the facts in (2.4) and (2.5), that is, $O(x_j,y_j) = p^{-1}(x_j,y_j)-1$ and $E(1-\delta|X) = E\{\delta O(X,Y)|X\}$.

It is not difficult to get that $E(J_1+J_2) = o(1)$. Denote $\widetilde{A} = E\{p^{-1}(S)XX^{\tau}\varepsilon^2\}+E\Big[\{p^{-1}(S)-1\}XX^{\tau}E^2(\varepsilon|X,\delta=0)\Big]$. We can have

$$\begin{aligned}
\mathrm{Cov}(J_1+J_2) &= \widetilde{A}+2E\left[\frac{\delta}{p(X,Y)}\left\{1-\frac{\delta}{p(X,Y)}\right\}XX^{\tau}\varepsilon E(\varepsilon|X,\delta=0)\right]\\
&= \widetilde{A}+2E\left[\{1-p^{-1}(X,Y)\}XX^{\tau}E^2(\varepsilon|X,\delta=0)\right] = A.
\end{aligned}$$

It follows that $n^{-1/2}\sum_{i=1}^{n}z_{i,W}(\beta)\xrightarrow{D}N(0,A)$. Further, we can have:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}z_{i,W}(\beta)z_{i,W}^{\tau}(\beta) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i^2}{p^2(x_i,y_i)}x_ix_i^{\tau}\varepsilon_i^2+\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\delta_i^2}{\hat{p}^2(x_i,y_i)}-\frac{\delta_i^2}{p^2(x_i,y_i)}\right\}x_ix_i^{\tau}\varepsilon_i^2\\
&:= J_1^{\star}+J_2^{\star}.
\end{aligned}$$

With regard to the term $J_2^{\star}$, it can be verified that

$$\begin{aligned}
|J_2^{\star}| &\le \frac{1}{n}\sum_{i=1}^{n}\sup\left\{\left|\frac{\delta_i^2}{\hat{p}^2(x_i,y_i)}-\frac{\delta_i^2}{p^2(x_i,y_i)}\right||x_ix_i^{\tau}|\varepsilon_i^2\right\}\\
&= \frac{1}{n}\sum_{i=1}^{n}|x_ix_i^{\tau}|\varepsilon_i^2\times\sup\left\{\left|\frac{\delta_i^2}{\hat{p}^2(x_i,y_i)}-\frac{\delta_i^2}{p^2(x_i,y_i)}\right|\right\}.
\end{aligned}$$

For the term $n^{-1}\sum_{i=1}^{n}|x_ix_i^{\tau}|\varepsilon_i^2$, we can have $n^{-1}\sum_{i=1}^{n}|x_ix_i^{\tau}|\varepsilon_i^2 = E(|XX^{\tau}|\varepsilon^2)+o_p(1)$ by the law of large numbers. While $\sup|\frac{\delta_i^2}{\hat{p}^2(x_i,y_i)}-\frac{\delta_i^2}{p^2(x_i,y_i)}| = o_p(1)$ since $\hat{p}$ is a consistent estimate of $p(x_i,y_i)$. Thus we can conclude that $J_2^{\star} = o_p(1)$. Thus, combining the above equations, we obtain $n^{-1}\sum_{i=1}^{n}z_{i,W}(\beta)z_{i,W}^{\tau}(\beta)\xrightarrow{P}B_1$, where $B_1 = E\{p^{-1}(S)XX^{\tau}\varepsilon^2\}$.

(b) Recall the expression of Formula (2.10) with some elementary calculations, we can obtain

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}z_{i,1}(\beta) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_i\Big[\{\delta_iy_i+(1-\delta_i)m_0(x_i)-x_i^{\tau}\beta\}+(1-\delta_i)\{\hat{m}_0(x_i)-m_0(x_i)\}\Big]\\
&:= J_3+J_4 \qquad\qquad (A.4)
\end{aligned}$$

where

$$J_3 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_i\varepsilon_i-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_i(1-\delta_i)\{y_i-m_0(x_i)\}. \qquad (A.5)$$

For the term J_4, similar to the argument for J_2, we have:

$$\begin{aligned}
J_4 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i(1-\delta_i)\frac{\sum\limits_{j=1}^{n}\delta_j\exp(\gamma y_j)\{y_j-m_0(x_i)\}\mathcal{K}_h(x_i,x_j)}{\sum\limits_{j=1}^{n}\delta_j\exp(\gamma y_j)\mathcal{K}_h(x_i,x_j)} \\
&= \frac{1}{n^{3/2}}\sum_{i=1}^{n} x_i(1-\delta_i)\frac{\sum\limits_{j=1}^{n}\delta_j\exp(\gamma y_j)\{y_j-m_0(x_i)\}\mathcal{K}_h(x_i,x_j)}{f(x_i)\{1-p(x_i)\}\alpha^{-1}(x_i;\gamma)} + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}[\delta_j\exp(\gamma y_j)\{y_j-m_0(x_j)\}]E\left[\frac{X(1-\delta)\mathcal{K}_h(X,x_j)}{f(X)\{1-p(X)\}\alpha^{-1}(X;\gamma)}\middle| x_j\right] + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\delta_j O(x_j,y_j)x_j\{y_j-m_0(x_j)\} + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\delta_j\{p^{-1}(x_j,y_j)-1\}x_j\{y_j-m_0(x_j)\} + o_p(1). \qquad \text{(A.6)}
\end{aligned}$$

Taking the expressions (A.5) and (A.6) into Eq. (A.4), we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,1}(\beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\varepsilon_i - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{1-\frac{\delta_i}{p(x_i,y_i)}\right\}x_i\{y_i-m_0(x_i)\} + o_p(1).$$

Note that $y_i = x_i^\tau\beta + \varepsilon_i$. It is easy to see that $m_0(x_i) = x_i^\tau\beta + E(\varepsilon|x_i,\delta=0)$. Thus we can obtain that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,1}(\beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\frac{\delta_i}{p(x_i,y_i)}\varepsilon_i + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{1-\frac{\delta_i}{p(x_i,y_i)}\right\}x_iE(\varepsilon|x_i,\delta=0) + o_p(1).$$

Thus, we can obtain $n^{-1/2}\sum_{i=1}^{n} z_{i,1}(\beta) \xrightarrow{D} N(0,A)$. Further,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} z_{i,1}(\beta)z_{i,1}^\tau(\beta) &= \frac{1}{n}\sum_{i=1}^{n} x_ix_i^\tau\{\delta_iy_i + (1-\delta_i)m_0(x_i) - x_i^\tau\beta\}^2 + o_p(1) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_ix_i^\tau\{\delta_i\varepsilon_i + (1-\delta_i)E(\varepsilon|x_i,\delta=0)\}^2 + o_p(1).
\end{aligned}$$

From the above expression, we obtain $n^{-1}\sum_{i=1}^{n} z_{i,1}(\beta)z_{i,1}^\tau(\beta) \xrightarrow{P} B_2$, where $B_2 = E\Big[p(S)XX^\tau\varepsilon^2 + \{1-p(S)\}XX^\tau E^2(\varepsilon|X,\delta=0)\Big]$.

(c) Invoking Formula (2.10) for $k=2$, we have

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,2}(\beta) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\left[\frac{\delta_i}{p(x_i,y_i)}y_i + \left\{1-\frac{\delta_i}{p(x_i,y_i)}\right\}m_0(x_i) - x_i^\tau\beta\right] \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\left\{\frac{\delta_i}{\hat{p}(x_i,y_i)} - \frac{\delta_i}{p(x_i,y_i)}\right\}\{y_i-m_0(x_i)\} \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\left\{1-\frac{\delta_i}{\hat{p}(x_i,y_i)}\right\}\{\hat{m}_0(x_i)-m_0(x_i)\} \\
&:= J_5 + J_6 + J_7.
\end{aligned}$$

We now deal with J_5, J_6 and J_7 separately. Using the same argument for J_2, we have

$$\begin{aligned}
J_6 &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\{(1-\delta_j)-\delta_jO(z_j)\}E\left[\frac{\delta X\{Y-m_0(X)\}O(X,Y)\mathcal{K}_h(X,x_j)}{f(X)\{1-p(X)\}}\middle| x_j\right] + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\{(1-\delta_j)-\delta_jO(z_j)\}x_j\frac{E[(1-\delta)\{Y-m_0(X)\}|x_j]}{E(1-\delta|x_j)} + o_p(1) = o_p(1).
\end{aligned}$$

The last equation follows from the fact that $E[(1-\delta)\{Y-m_0(X)\}|x_j]/E(1-\delta|x_j) = E\{Y-m_0(X)|x_j, \delta=0\} \equiv 0$. Similarly, we can prove $J_7 = o_p(1)$. Further, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,2}(\beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\Big[\frac{\delta_i}{p(x_i,y_i)}\varepsilon_i + \Big\{1-\frac{\delta_i}{p(x_i,y_i)}\Big\}E(\varepsilon|x_i,\delta=0)\Big] + o_p(1).$$

Note that the asymptotic expansion of the above term is the same as that for $n^{-1/2}\sum_{i=1}^{n} z_{i,1}(\beta)$. The application of the Central Limit Theorem yields that $n^{-1/2}\sum_{i=1}^{n} z_{i,2}(\beta) \xrightarrow{D} N(0,A)$. Furthermore,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} z_{i,2}(\beta)z_{i,2}^{\tau}(\beta) &= \frac{1}{n}\sum_{i=1}^{n} x_i x_i^{\tau}\Big[\frac{\delta_i}{p(x_i,y_i)}y_i + \Big\{1-\frac{\delta_i}{p(x_i,y_i)}\Big\}m_0(x_i) - x_i^{\tau}\beta\Big]^2 + o_p(1)\\ &= \frac{1}{n}\sum_{i=1}^{n} x_i x_i^{\tau}\Big[\frac{\delta_i}{p(x_i,y_i)}\varepsilon_i + \Big\{1-\frac{\delta_i}{p(x_i,y_i)}\Big\}E(\varepsilon|X,\delta=0)\Big]^2 + o_p(1).\end{aligned}$$

Thus, $n^{-1}\sum_{i=1}^{n} z_{i,2}(\beta)z_{i,2}^{\tau}(\beta) \xrightarrow{P} A$, which completes the proof of Lemma 1. Note that $\hat{\theta}_{B_1,i} = z_{i,W}(\beta) - \hat{p}^{-1}(s_i)\delta_i x_i x_i^{\tau}(\hat{\beta}-\beta)$. Together the result that $n^{-1}\sum_{i=1}^{n} z_{i,W}(\beta)z_{i,W}^{\tau}(\beta) \xrightarrow{P} B_1$ with the fact that $\hat{\beta}-\beta = O_p(n^{-1/2})$, we can easily obtain the consistency of the estimate of B_1. The consistencies of $\hat{B}_2$ and $\hat{A}$ can be similarly obtained and we omit the details here. □

Lemma 2. *Under conditions* C1–C6 *in the Appendix, if β is the true parameter, then*

(i) *If $\hat{\gamma}$ is calculated from an independent survey, we have*

$$ⓐ\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,W}(\beta) \xrightarrow{D} N(0,V), \qquad \frac{1}{n}\sum_{i=1}^{n} z_{i,W}(\beta)z_{i,W}^{\tau}(\beta) \xrightarrow{P} B_1, \tag{A.7}$$

$$ⓑ\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,1}(\beta) \xrightarrow{D} N(0,V), \qquad \frac{1}{n}\sum_{i=1}^{n} z_{i,1}(\beta)z_{i,1}^{\tau}(\beta) \xrightarrow{P} B_2, \tag{A.8}$$

$$ⓒ\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,2}(\beta) \xrightarrow{D} N(0,V), \qquad \frac{1}{n}\sum_{i=1}^{n} z_{i,2}(\beta)z_{i,2}^{\tau}(\beta) \xrightarrow{P} A. \tag{A.9}$$

(ii) *If $\hat{\gamma}$ is computed from a validation sample, we have*

$$ⓐ\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,W}(\beta) \xrightarrow{D} N(0,\widetilde{V}), \qquad \frac{1}{n}\sum_{i=1}^{n} z_{i,W}(\beta)z_{i,W}^{\tau}(\beta) \xrightarrow{P} B_1, \tag{A.10}$$

$$ⓑ\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,1}(\beta) \xrightarrow{D} N(0,\widetilde{V}), \qquad \frac{1}{n}\sum_{i=1}^{n} z_{i,1}(\beta)z_{i,1}^{\tau}(\beta) \xrightarrow{P} B_2, \tag{A.11}$$

$$ⓒ\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,2}(\beta) \xrightarrow{D} N(0,\widetilde{V}), \qquad \frac{1}{n}\sum_{i=1}^{n} z_{i,2}(\beta)z_{i,2}^{\tau}(\beta) \xrightarrow{P} A \tag{A.12}$$

where the above notations have been defined in Section 3.2 in detail.

Proof. Consider (i) ⓐ . Denote $\gamma^{\star}$ as the true value of the parameter γ. Using the same argument for Lemma 1, the following decomposition holds:

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,W}(\beta) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i}{\hat{p}(s_i,\gamma^{\star})}x_i\varepsilon_i + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\frac{\delta_i}{\hat{p}(s_i,\hat{\gamma})} - \frac{\delta_i}{\hat{p}(s_i,\gamma^{\star})}\Big\}x_i\varepsilon_i\\ &= J_1 + J_2 + \frac{1}{n}\sum_{i=1}^{n}\delta_i x_i\varepsilon_i\frac{\partial\hat{p}^{-1}(s_i,\gamma)}{\partial\gamma}\Big|_{\gamma=\gamma_1}\sqrt{n}(\hat{\gamma}-\gamma^{\star})\\ &:= J_1 + J_2 + \tilde{J}_1,\end{aligned} \tag{A.13}$$

here γ_1 is the line segment between $\hat{\gamma}$ and $\gamma^{\star}$. Recall $\hat{p}^{-1}(s_i,\gamma) = 1 + \hat{\alpha}(x_i;\gamma)\exp(\gamma y_i)$ and (2.6), (2.8). We can then derive that

$$\frac{\partial\hat{p}^{-1}(s_i,\gamma)}{\partial\gamma} = \hat{\alpha}(x_i;\gamma)\exp(\gamma y_i)\{y_i - \hat{m}_0(x_i;\gamma)\}. \tag{A.14}$$

Combining the expression $O(s_i) := O(x_i, y_i) = \exp(\gamma y_i)\alpha(x_i;\gamma)$, we can further obtain that

$$\begin{aligned}
\tilde{J}_1 &= \frac{1}{n}\sum_{i=1}^{n} \delta_i x_i \varepsilon_i \hat{\alpha}(x_i;\gamma_1)\exp(\gamma_1 y_i)\{y_i - \hat{m}_0(x_i;\gamma_1)\}\sqrt{n}(\hat{\gamma}-\gamma^\star) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i \delta_i O(s_i)\{y_i - m_0(x_i;\gamma^\star)\}\sqrt{n}(\hat{\gamma}-\gamma^\star) + o_p(1) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i (1-\delta_i)\{\varepsilon_i - E(\varepsilon|X,\delta=0)\}\sqrt{n}(\hat{\gamma}-\gamma^\star) + o_p(1) \\
&= E\Big[X\{1-p(S;\gamma)\}\{\varepsilon - E(\varepsilon|X,\delta=0)\}^2\Big]\sqrt{n}(\hat{\gamma}-\gamma^\star) + o_p(1)..
\end{aligned} \tag{A.15}$$

As a result, $\tilde{J}_1$ converges to $N(0, HH^\tau V_\gamma)$, where $H = E\Big[X\{1-p(S;\gamma)\}\{\varepsilon - E(\varepsilon|X,\delta=0)\}^2\Big]$. It then follows that $n^{-1/2}\sum_{i=1}^{n} z_{i,W}(\beta) \xrightarrow{D} N(0,V)$, where $V = A + HH^\tau V_\gamma$.

Consider (i) ⓑ. Now we turn to consider the term $z_{i,1}(\beta)$. Similarly

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,1}(\beta) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\Big[\{\delta_i y_i + (1-\delta_i)\hat{m}_0(x_i;\gamma^\star) - x_i^\tau\beta\} + (1-\delta_i)\{\hat{m}_0(x_i;\hat{\gamma}) - \hat{m}_0(x_i;\gamma^\star)\}\Big] \\
&= J_3 + J_4 + \frac{1}{n}\sum_{i=1}^{n} x_i\,(1-\delta_i)\frac{\partial \hat{m}_0(x_i;\gamma)}{\partial\gamma}\bigg|_{\gamma=\gamma_2}\sqrt{n}(\hat{\gamma}-\gamma^\star) \\
&:= J_3 + J_4 + \tilde{J}_3,
\end{aligned} \tag{A.16}$$

where γ_2 is the value between $\hat{\gamma}$ and $\gamma^\star$. Note that the expression of $\hat{m}_0(x_i;\gamma)$ has been defined in (2.8), then

$$\frac{\partial \hat{m}_0(x_i;\gamma)}{\partial\gamma} = \frac{\sum_{j=1}^{n}\delta_j\exp(\gamma y_j)y_j^2\mathcal{K}_h(x_i,x_j)}{\sum_{j=1}^{n}\delta_j\exp(\gamma y_j)\mathcal{K}_h(x_i,x_j)} - \hat{m}_0^2(x_i;\gamma) = \hat{E}(Y^2|x_i,\delta=0) - \hat{m}_0^2(x_i;\gamma). \tag{A.17}$$

Thus, we have

$$\begin{aligned}
\tilde{J}_3 &= \frac{1}{n}\sum_{i=1}^{n} x_i(1-\delta_i)\{E(Y^2|x_i,\delta=0) - m_0^2(x_i;\gamma)\}\sqrt{n}(\hat{\gamma}-\gamma^\star) + o_p(1) \\
&= E\Big[X(1-\delta_i)\{Y - m_0(X;\gamma)\}^2\Big]\sqrt{n}(\hat{\gamma}-\gamma^\star) + o_p(1) \\
&= E\Big[X\{1-p(S;\gamma)\}\{\varepsilon - E(\varepsilon|X,\delta=0)\}^2\Big]\sqrt{n}(\hat{\gamma}-\gamma^\star) + o_p(1).
\end{aligned}$$

This yields that $\tilde{J}_3$ converges to $N(0, HH^\tau V_\gamma)$. It then follows $n^{-1/2}\sum_{i=1}^{n} z_{i,1}(\beta) \xrightarrow{D} N(0,V)$.

Deal with (i) ⓒ. Denote $\tilde{J}_5(s_i;\gamma) = \delta_i y_i/\hat{p}(s_i;\gamma) + \{1 - \delta_i/\hat{p}(s_i;\gamma)\}\hat{m}_0(x_i;\gamma) - x_i^\tau\beta,\ i = 1,\ldots,n$. For the auxiliary random vector $z_{i,2}$, we have

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,2}(\beta) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\Big[\frac{\delta_i}{\hat{p}(s_i,\hat{\gamma})}y_i + \Big\{1 - \frac{\delta_i}{\hat{p}(s_i,\hat{\gamma})}\Big\}\hat{m}_0(x_i;\hat{\gamma}) - x_i^\tau\beta\Big] \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i[\tilde{J}_5(s_i;\gamma^\star) + \{\tilde{J}_5(s_i;\hat{\gamma}) - \tilde{J}_5(s_i;\gamma^\star)\}] \\
&= J_5 + J_6 + J_7 + \frac{1}{n}\sum_{i=1}^{n} x_i\frac{\partial\tilde{J}_5(s_i;\gamma)}{\partial\gamma}\bigg|_{\gamma=\gamma_3}\sqrt{n}(\hat{\gamma}-\gamma^\star) \\
&= J_5 + J_6 + J_7 + \tilde{J}_6,
\end{aligned}$$

where γ_3 is one value between $\hat{\gamma}$ and the true value $\gamma^\star$. Recalling Formulae (A.14) and (A.17), we can gain

$$\begin{aligned}
\frac{\partial\tilde{J}_6(s_i;\gamma)}{\partial\gamma} &= \delta_i\{y_i - \hat{m}_0(x_i;\hat{\gamma})\}\frac{\partial\hat{p}^{-1}(s_i,\gamma)}{\partial\gamma} + \Big\{1 - \frac{\delta_i}{\hat{p}(s_i,\gamma)}\Big\}\frac{\partial\hat{m}_0(x_i;\gamma)}{\partial\gamma} \\
&= \delta_i\hat{\alpha}(x_i;\gamma)\exp(\gamma y_i)\{y_i - \hat{m}_0(x_i;\hat{\gamma})\}^2 + \Big\{1 - \frac{\delta_i}{\hat{p}(s_i,\gamma)}\Big\}\{\hat{E}(Y^2|x_i,\delta=0) - \hat{m}_0^2(x_i;\gamma)\}.
\end{aligned}$$

This leads to

$$\begin{aligned}
\tilde{J}_6 &= \frac{1}{n}\sum_{i=1}^{n} x_i\Big[\delta_i O(s_i,\gamma)\{y_i - m_0(x_i;\gamma)\}^2 + \Big\{1 - \frac{\delta_i}{\hat{p}(s_i,\gamma)}\Big\} \\
&\quad \times \{E(Y^2|x_i,\delta=0) - \hat{m}_0^2(x_i;\gamma)\}\Big]\sqrt{n}(\hat{\gamma}-\gamma^\star) + o_p(1) \\
&= E\Big[X(1-\delta)\{Y - m_0(X;\gamma)\}^2\Big]\sqrt{n}(\hat{\gamma}-\gamma^\star) + o_p(1) \\
&= E\Big[X\{1-p(S;\gamma)\}\{\varepsilon - E(\varepsilon|X,\delta=0)\}^2\Big]\sqrt{n}(\hat{\gamma}-\gamma^\star) + o_p(1).
\end{aligned}$$

As a result, we can derive that $n^{-1/2}\sum_{i=1}^{n} z_{i,2}(\beta) \xrightarrow{D} N(0,V)$.

We are now in the position to prove the results in (ii) with the estimate $\hat{\gamma}$ being acquired from a validation sample.

Consider (ii) ⓐ . Similarly, for $z_{i,W}(\beta)$, the same arguments with Formula (A.13) yield

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,W}(\beta) &= J_1 + J_2 + \frac{1}{n}\sum_{i=1}^{n}\delta_i x_i \varepsilon_i \frac{\partial \hat{p}^{-1}(s_i,\gamma)}{\partial\gamma}\Big|_{\gamma=\gamma_4}\sqrt{n}(\hat{\gamma}-\gamma^\star) \\
&:= J_1 + J_2 + J_{v1} \qquad \text{(A.18)}
\end{aligned}$$

where γ_4 is the value between $\hat{\gamma}$ and $\gamma^\star$. Also, by Eq. (3.1), it follows that

$$\begin{aligned}
0 &= \sum_{i=1}^{n}(1-\delta_i)r_i\{y_i - \hat{m}_0(x_i,\hat{\gamma})\} \\
&= \sum_{i=1}^{n}(1-\delta_i)r_i\{y_i - \hat{m}_0(x_i,\gamma_0)\} - \sum_{i=1}^{n}(1-\delta_i)r_i\frac{\partial \hat{m}_0(x_i,\gamma)}{\partial\gamma}\Big|_{\gamma=\gamma_0}(\hat{\gamma}-\gamma_0),
\end{aligned}$$

here, γ_0 is the probability limit of $\hat{\gamma}$ and $\gamma_0 = \gamma^\star$. Combining Formula (A.17) with the above equation, we can derive that

$$\begin{aligned}
\sqrt{n}(\hat{\gamma}-\gamma^\star) &= \Big\{\frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)r_i\frac{\partial \hat{m}_0(x_i,\gamma)}{\partial\gamma}\Big|_{\gamma=\gamma^\star}\Big\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)r_i\{y_i - \hat{m}_0(x_i,\gamma^\star)\} \\
&:= \Big(E\Big[r(1-\delta)\{E(Y^2|x_i,\delta=0) - m_0^2(x_i;\gamma^\star)\}\Big]\Big)^{-1} \\
&\quad \times \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)r_i\Big[\{y_i - m_0(x_i,\gamma^\star)\} + \{m_0(x_i,\gamma^\star) - \hat{m}_0(x_i,\gamma^\star)\}\Big], \qquad \text{(A.19)}
\end{aligned}$$

where,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)r_i\{m_0(x_i,\gamma^\star) - \hat{m}_0(x_i,\gamma^\star)\} \\
&= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\delta_i \exp(\gamma^\star y_i)\{y_i - m_0(x_i;\gamma^\star)\}\Big]E\Big[\frac{r(1-\delta)\mathcal{K}_h(X,x_i)}{f(X)\{1-p(X)\}\alpha^{-1}(X;\gamma)}\Big| x_j\Big] + o_p(1) \\
&= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_i \nu\{p^{-1}(x_i,y_i) - 1\}\{y_i - m_0(x_i;\gamma^\star)\} + o_p(1), \qquad \text{(A.20)}
\end{aligned}$$

here $\nu = E(r|\delta=0)$ and $f(\cdot)$ is the marginal density of X. Similar to the derivation of $\tilde{J}_1$ in (A.15), taking Formulae (A.19) and (A.20) into the term J_{v1} in (A.18), we can derive

$$\begin{aligned}
J_{v1} &= E\Big[X\{1-p(S;\gamma)\}\{\varepsilon - E(\varepsilon|X,\delta=0)\}^2\Big]\sqrt{n}(\hat{\gamma}-\gamma^\star) + o_p(1) \\
&= E\Big[X\{1-p(S;\gamma)\}\{\varepsilon - E(\varepsilon|X,\delta=0)\}^2\Big]\Big(E\Big[r(1-\delta)\Big\{E(Y^2|x_i,\delta=0) - m_0^2(x_i;\gamma^\star)\Big\}\Big]\Big)^{-1} \\
&\quad \times \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[(1-\delta_i)r_i - \delta_i\nu\{p^{-1}(x_i,y_i) - 1\}\Big]\{y_i - m_0(x_i,\gamma^\star)\} \\
&= HM^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[(1-\delta_i)r_i - \delta_i\nu\{p^{-1}(x_i,y_i) - 1\}\Big]\{y_i - m_0(x_i,\gamma^\star)\}.
\end{aligned}$$

Recalling the expression of $n^{-1/2}\sum_{i=1}^{n} z_{i,W}(\beta)$ in (A.18), we derive that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,W}(\beta) \sim N(0, \widetilde{V}),$$

where $\widetilde{V} = \mathrm{Var}(\eta_1)$ and

$$\eta_1 = \frac{\delta}{p(S)}X\varepsilon + X\left\{1 - \frac{\delta}{p(S)}\right\}E(\varepsilon|X, \delta = 0) + HM^{-1}\Big[(1-\delta)r - \delta\nu\{p^{-1}(S) - 1\}\Big]\{Y - m_0(X, \gamma^\star)\}$$

with $H = E\Big[X\{1 - p(S;\gamma)\}\{\varepsilon - E(\varepsilon|X, \delta = 0)\}^2\Big]$ and $M = E\Big[r(1-\delta)\Big\{E(Y^2|x_i, \delta = 0) - m_0^2(x_i;\gamma^\star)\Big\}\Big]$.

Deal with (ii) ⓑ. Referencing to the derivation of (A.16), we can obtain

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,1}(\beta) &= J_3 + J_4 + \frac{1}{n}\sum_{i=1}^{n} x_i(1-\delta_i)\frac{\partial \hat{m}_0(x_i;\gamma)}{\partial\gamma}\bigg|_{\gamma=\gamma_5}\sqrt{n}(\hat{\gamma} - \gamma^\star) \\ &:= J_3 + J_4 + \tilde{J}_{v2}.\end{aligned}$$

Similar to the derivation of (A.18), it is easy to see that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,1}(\beta) \sim N(0, \widetilde{V}).$$

Consider (ii) ⓒ. We can similarly derive that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{i,2}(\beta) \sim N(0, \widetilde{V}).$$

Since $\hat{p}(s_i)$ and $\hat{m}_0(x_i)$ are the consistent estimates that are $\hat{p}(s_i) = p(x_i, y_i) + o_p(1)$ and $\hat{m}_0(x_i) = m_0(x_i) + o_p(1)$, by the Law of Large Numbers and the continuous mapping theorem, we can derive that

$$\frac{1}{n}\sum_{i=1}^{n} z_i(\beta) z_i^\tau(\beta) \xrightarrow{P} B,$$

where $z_i(\beta)$ in (A.8) can be also replaced by $z_{i,W}(\beta)$ or $z_{i,k}(\beta),\ k = 1, 2$, and $B := B_1 = E\Big\{p^{-1}(S)XX^\tau\varepsilon^2\Big\}$, $B := B_2 = E\Big[p(S)XX^\tau\varepsilon^2 + \{1 - p(S)\}XX^\tau E^2(\varepsilon|X, \delta = 0)\Big]$ and $B := A = E\Big[p^{-1}(S)XX^\tau\varepsilon^2 + \{1 - p^{-1}(S)\}XX^\tau E^2(\varepsilon|X, \delta = 0)\Big]$, respectively. □

Proof of Theorem 1. Consider the known $\gamma^\star$ case. Recalling the description in Section 2.1, $\hat{\beta}_W$ is the weighted empirical likelihood estimate. Denote

$$T_{1n}(\beta, \lambda_W) = \frac{1}{n}\sum_{i=1}^{n}\frac{z_{i,W}}{1 + \lambda_W^\tau z_{i,W}},$$

$$T_{2n}(\beta, \lambda_W) = \frac{\partial\{-l_W(\beta)\}}{\partial\beta} = \frac{1}{n}\sum_{i=1}^{n}\frac{\lambda_W z_{i,W}'}{1 + \lambda_W^\tau z_{i,W}},$$

where $z_{i,W}' = \partial z_{i,W}/\partial\beta = -\delta_i x_i x_i^\tau/\hat{p}(z_i)$. Thus, $\hat{\beta}_W$ and $\tilde{\lambda}_W$ are the roots of the following two equations $T_{1n} = 0$ and $T_{2n} = 0$. Applying Taylor expansions to $T_{1n}(\hat{\beta}_W, \tilde{\lambda}_W)$ and $T_{2n}(\hat{\beta}_W, \tilde{\lambda}_W)$ at $(\beta, 0)$, we can get

$$0 = T_{1n}(\hat{\beta}_W, \tilde{\lambda}_W) = T_{1n}(\beta, 0) + \frac{\partial T_{1n}(\beta, 0)}{\partial\beta}(\hat{\beta}_W - \beta) + \frac{\partial T_{1n}(\beta, 0)}{\partial\lambda}\tilde{\lambda}_W + o_p(u_n),$$

$$0 = T_{2n}(\hat{\beta}_W, \tilde{\lambda}_W) = T_{2n}(\beta, 0) + \frac{\partial T_{2n}(\beta, 0)}{\partial\beta}(\hat{\beta}_W - \beta) + \frac{\partial T_{2n}(\beta, 0)}{\partial\lambda}\tilde{\lambda}_W + o_p(u_n),$$

where $u_n = \|\hat{\beta}_W - \beta\| + \|\tilde{\lambda}_W\|$. Note that $T_{1n}(\beta, 0) = \sum_{i=1}^{n} z_{i,W}/n$ and $T_{2n}(\beta, 0) = 0$. Then the above two equations can be rewritten as

$$\begin{bmatrix}\tilde{\lambda}_W \\ \hat{\beta}_W - \beta\end{bmatrix} = M_n^{-1}\begin{bmatrix}-\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} z_{i,W} + o_p(u_n) \\ o_p(u_n)\end{bmatrix},$$

where

$$M_n = \begin{bmatrix} \dfrac{\partial T_{1n}(\beta, \lambda_W)}{\partial \lambda_W} & \dfrac{\partial T_{1n}(\beta, \lambda_W)}{\partial \beta} \\ \dfrac{\partial T_{2n}(\beta, \lambda_W)}{\partial \lambda_W} & \dfrac{\partial T_{2n}(\beta, \lambda_W)}{\partial \beta} \end{bmatrix}_{(\beta,\lambda_W)=(\beta,0)}$$

$$= \begin{bmatrix} -\dfrac{1}{n}\sum_{i=1}^{n} z_{i,W} z_{i,W}^{\tau} & \dfrac{1}{n}\sum_{i=1}^{n} z'_{i,W} \\ \dfrac{1}{n}\sum_{i=1}^{n} z'_{i,W} & 0 \end{bmatrix}.$$

Therefore, we can obtain

$$\sqrt{n}(\hat{\beta}_W - \beta) = \Big(-\frac{1}{n}\sum_{i=1}^{n} z'_{i,W}\Big)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_{i,W} + o_p(1)$$

$$= \Big(\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i x_i x_i^{\tau}}{\hat{p}(z_i)}\Big)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_{i,W} + o_p(1).$$

Together with the result (A.1) in Lemma 1, we can derive that

$$\sqrt{n}(\hat{\beta}_W - \beta) \xrightarrow{D} N(0, T^{-1}AT^{-1}),$$

where $T = E(XX^{\tau})$.

From (A.21), we know

$$\hat{l}_W(\beta, \gamma^{\star}) = 2\sum_{i=1}^{n} \log(1 + \lambda_W^{\tau} \hat{z}_{i,W}). \tag{A.21}$$

Together the proof in Theorem 1 of Xue (2009a,b) with that for Lemma 1, we get

$$\hat{l}_W(\beta, \gamma^{\star}) = \Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{z}_{i,W}\Big)^{\tau} \Big(\frac{1}{n}\sum_{i=1}^{n} \hat{z}_{i,W}\hat{z}_{i,W}^{\tau}\Big)^{-1} \Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{z}_{i,W}\Big) + o_p(1).$$

Lemma 1 yields the result.

In the same way, denote

$$\hat{l}_{I,k}(\beta, \gamma^{\star}) = 2\sum_{i=1}^{n} \log(1 + \lambda_{I,k}^{\tau} \hat{z}_{i,k}), \quad k = 1, 2$$

we can conclude that $\sqrt{n}(\hat{\beta}_{I,k} - \beta) \xrightarrow{D} N(0, T^{-1}AT^{-1}),\ k = 1, 2.$ □

Proof of Theorem 2. By Lemma 2(i), similar to the proof of Theorem 1, we can prove Theorem 2. □

Proof of Theorem 3. Similar to the proof of Theorem 1, together with Lemma 2(ii), we can conclude the proof of Theorem 3. □

References

Hall, P., La Scala, B., 1990. Methodology and algorithms of empirical likelihood. Internat. Statist. Rev. 58, 109–127.

Imai, K., 2009. Statistical analysis of randomized experiments with non-ignorable missing binary outcomes: an application to a voting experiment. J. R. Stat. Soc. Ser. C 58, 83–104.

Jamshidian, M., Yuan, K.H., 2013. Data-driven sensitivity analysis to detect missing data mechanism with applications to structural equation modelling. J. Stat. Comput. Simul. 83, 1344–1362.

Kim, J.K., Yu, C.L., 2011. A semiparametric estimation of mean functionals with nonignorable missing data. J. Amer. Statist. Assoc. 106, 157–165.

Köhler, M., Schindler, A., Sperlich, S., 2014. A review and comparison of bandwidth selection methods for kernel regression. Internat. Statist. Rev. (in press).

Lee, S.Y., Tang, N.S., 2006. Analysis of nonlinear structural equation models with nonignorable missing covariates and ordered categorical data. Statist. Sinica 16, 1117–1141.

Lipsitz, S.R., Molenberghs, G., Fitzmaurice, G.M., Ibrahim, J.G., 2004. Protective estimate for linear regression with nonignorably missing Gaussian outcomes. Stat. Model. 4, 3–17.

Noémie, R., Roch, G., Xavier, P., 2011. Sensitivity analysis when data are missing not-at-random. Epidemiology 22, 282–283.

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

Owen, A.B., 1990. Empirical likelihood ratio confidence regions. Ann. Statist. 18, 90–120.

Owen, A.B., 1991. Empirical likelihood for linear models. Ann. Statist. 19, 1725–1747.

Qin, Y.S., Lei, Q.Z., 2010. On empirical likelihood for linear models with missing responses. J. Statist. Plann. Inference 140, 3399–3408.

Qin, J., Zhang, B., Leung, D.H.Y., 2009. Empirical likelihood in missing data problems. J. Amer. Statist. Assoc. 104, 1492–1503.

Rotnitzky, A., Robins, J., Scharfstein, D., 1998. Semiparametric regression for repeated outcomes with non-ignorable non-response. J. Amer. Statist. Assoc. 93, 1321–1339.

Sinha, S.K., 2012. Robust analysis of longitudinal data with nonignorable missing responses. Metrika 75, 913–938.

Tang, N.S., Zhao, P.Y., 2013. Empirical likelihood-based inference in nonlinear regression models with missing responses at random. Statistics 47, 1141–1159.

Ware, J.H., Dockery, D.W., Spiro III, A., Speizer, F.E., Ferris Jr., B.G., 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. Am. Rev. Respir. Dis. 129, 366–374.

Xie, H., Qian, Y., Qu, L.M., 2011. A semiparametric approach for analyzing nonignorable missing data. Statist. Sinica 21, 1881–1899.

Xue, L.G., 2009a. Empirical likelihood for linear models with missing responses. J. Multivariate Anal. 100, 1353–1366.

Xue, L.G., 2009b. Empirical likelihood confidence intervals for response mean with data missing at random. Scand. J. Statist. 36, 671–685.

Xue, L.G., Xue, D., 2011. Empirical likelihood for semiparametric regression model with missing response data. J. Multivariate Anal. 102, 723–740.

Zhao, H., Zhao, P.Y., Tang, N.S., 2013. Empirical likelihood inference for mean functionals with nonignorably missing response data. Comput. Statist. Data Anal. 66, 101–116.

Zhu, L.X., Xue, L.G., 2006. Empirical likelihood confidence regions in a partially linear single-index model. J. R. Stat. Soc. Ser. B 68, 549–570.