Statistica Sinica Preprint No: SS-2016-0388R1	
Title	Rank-based estimating equation with non-ignorable missing responses via empirical likelihood
Manuscript ID	SS-2016-0388R1
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0388
Complete List of Authors	Huybrechts F. Bindele and Yichuan Zhao
Corresponding Author	Huybrechts F. Bindele
E-mail	hbindele@southalabama.edu
Notice: Accepted version subject to English editing.	

RANK-BASED ESTIMATING EQUATION WITH NON-IGNORABLE MISSING RESPONSES VIA EMPIRICAL LIKELIHOOD

Huybrechts F. Bindele[1] and Yichuan Zhao[2]

[1]*University of South Alabama and* [2]*Georgia State University*

Abstract: In this paper, a general regression model with responses missing not at random is considered. From a rank-based estimating equation, a rank-based estimator of the regression parameter is derived. Based on this estimator's asymptotic normality property, a consistent sandwich estimator of its corresponding asymptotic covariance matrix is obtained. In order to overcome the over-coverage issue of the normal approximation procedure, the empirical likelihood based on the rank-based gradient function is defined, and its asymptotic distribution is established. Extensive simulation experiments under different settings of error distributions with different response probabilities are considered, and the simulation results show that the proposed empirical likelihood approach has better performance in terms of coverage probability and average length of confidence intervals for the regression parameters compared with the normal approximation approach and its least-squares counterpart. A real data example is provided to illustrate the proposed methods.

Key words and phrases: Empirical likelihood, Imputation, Non-ignorable missing, Rank-based estimator.

[1] *Corresponding author address*: 411 University Blvd. N, ILB 316, Dept. of Mathematics and Statistics, University of South Alabama, Mobile AL 36688-0002. Email - hbindele@southalabama.edu

[2] *yichuan@gsu.edu*

1. Introduction:

Missing data have become unavoidable in the statistical community and have garnered a lot of attention within the last few decades. The missingness occurrence is subject to a number of common reasons, including equipment malfunction, contamination of samples, manufacturing defects, drop out in clinical trials, weather conditions, and incorrect data entry. For missing data problems, the missing mechanism often encountered is known as *missing at random* (MAR). This assumption asserts that the response probability can only depend on the values of those other variables that have been observed. It is a common assumption for statistical analysis in the presence of missing data and has been determined to be reasonable in many practical situations. Nevertheless, situations exist where the missingness of a response depends on the value of the unobserved outcome even after controlling for the covariates. In this case, statistical analysis with the MAR assumption would be invalid. To that end, this paper is concerned with the statistical inference of the true parameters in a regression model from which responses are subject to missingness. The missing data mechanism being considered is the *missing not at random* (MNAR) assumption discussed in Rubin (1976). Under this assumption, the probability that a response variable is missing depends on itself after controlling the predictors. As pointed out in Kim & Yu (2011), the MNAR condition exists, for example, in surveys of income, when the nonresponse rates tend to be higher for low socio-economic groups. This missingness type can be encountered in many fields of study such as social sciences, biomedical studies, agriculture economics and psychology among others. As an example of data, one may be interested in a survey of families in a city that includes many socioeconomic variables and a follow-up survey a few months or years later for the recording of new observations. By the time of the new recording, some families may have left the city, died, or can not be located thereby resulting in missing observations. Also in a clinical trial study, some patients may decide to drop out during the course of study which leads to some missing information. Many other scenarios can be found in the literature in the framework of missing data. To be more precise, consider the general regression model

$$y_i = g(\mathbf{x}_i, \boldsymbol{\beta}_0) + \varepsilon_i, \qquad 1 \le i \le n, \tag{1.1}$$

where $g : \mathbb{R}^p \times \mathscr{B} \to \mathbb{R}$ is fully specified, and $\boldsymbol{\beta}_0 \in \mathscr{B} \subset \mathbb{R}^p$ is a vector of parameters with $\mathscr{B}$ compact, $\mathbf{x}_i$'s are i.i.d. p-variable random covariate vectors, and conditional on $\mathbf{x}_i$, the model errors ε_i are continuous, i.i.d. with cumulative distribution F and corresponding density f, and $E(\varepsilon^2|\mathbf{x}) > 0$. Our interest is in making inference about the true value $\boldsymbol{\beta}_0$, when there are some responses missing in the general regression model (1.1).

There is a plenitude of literature on handling model (1.1) in the complete case analysis, that is, ignoring observations with missing responses. Such approaches include the least squares (LS), the least absolute deviation (LAD), and the maximum likelihood (ML) among others. In the statistical community, it is well understood that statistical inference based on the LS approach is efficient when the model assumptions such as normality of the model error distribution and homogeneity (constant variance of the model errors) are satisfied. The later assumption could be relaxed by considering the weighted least squares approach, but still assuming a normality of the model error distribution. Under the violation of these assumptions, the LS could lead to misleading inference. In an effort to come out with a robust approach, the LAD was proposed, but shown to be very less efficient in many situations. The ML approach on the other hand, is a very powerful alternative to the LS and LAD but requires model distribution specification. One of the main disadvantages of this approach is that in real life situations, mainly with missing data, it is almost impossible to specify the model error distribution. Under the MAR assumption however, as pointed out in Little & Rubin (2002), the complete case analysis, that is, ignoring observations with missing response, will lead to an efficient ML estimator. It is worth pointing out that even for the complete case analysis, the rank-based approach introduced by Jaeckel (1972) outperforms the aforementioned approaches in terms of robustness and efficiency when dealing with

heavy-tailed model errors and/or in the presence of outliers; see Hettmansperger & McKean (2011) for linear models and Bindele & Abebe (2012) for nonlinear models. Recently, for missing responses under the MAR assumption, a rank-based approach has been proposed by **?** for model (1.1), and by Bindele & Abebe (2015) for the semiparametric linear model.

Estimation under nonignorable missing responses is a very challenging problem and has captured a lot of attention in the last decade. The difficulty of this problem relies on the fact that the mechanism causing missingness is unknown, and both the response probability and the regression parameters need to be estimated. Some contributions that have attempted to solve this problem include those of Greenlees et al. (1982), Baker & Laird (1988), Chambers & Welsh (1993), Diggle & Kenward (1994), Ibrahim et al. (1999), and Ibrahim et al. (2001). These works provide an estimation of parameters under nonignorable missing data based on the maximum likelihood approach. A review of some parametric approaches for handling nonignorable missing data can be found in Molenberghs & Kenward (2007). Motivated by the work of Rotnitzky et al. (1998), Kim & Yu (2011) proposed an estimation procedure, where the response mechanism is modeled using the logistic semi-parametric regression model. Another issue that arises when considering regression models with nonignorable missing data is model identifiability. Identification of graphical models for nonignorable nonresponse of binary outcomes in longitudinal was investigated by Ma et al. (2003). Wang et al. (2014) proposed an instrumental variable approach for model identification and estimation, and more recently, Miao et al. (2016) proposed the identifiability of normal and normal mixture models with nonignorable missing data. Other recent developments for estimation approaches under nonignorable missing data include those of Zhao & Shao (2015), Shao & Wang (2016), Tang et al. (2016) and Fang et al. (2016).

One notes that most of the approaches discussed above are based on the normal approximation as a way to handle statistical inference. Unfortunately, normal approximation approaches require estimating the estimator's covariance matrix. While estimating the estimator's covariance matrix for the LS approach may not seem so difficult, this has been shown not to be a simple task when considering the rank-based objective function, mainly when dealing with dependent residuals; see Brunner & Denker (1994). The empirical likelihood (EL) approach on the other hand, is a way of avoiding estimating such a covariance matrix, conducting a direct inference about the true regression parameters, and overcoming the drawback of the normal approximation method (Owen, 1988, 1990). Qin & Lawless (1994) developed the EL inference procedure for general estimating equations for complete data, and Owen (2001) makes an excellent summary about the theory and applications of the EL methods. Recent progress in the EL method includes linear transformation models with right censoring (Yu et al. (2011), Yang & Zhao (2012)), the jackknife EL procedure (Jing et al. (2009), Gong et al. (2010), Zhang & Zhao (2013), Yang & Zhao (2013), and Yang & Zhao (2015)), high dimensional EL method (Chen et al. (2009), Hjort et al. (2009), Tang & Leng (2010), Lahiri et al. (2012)), and the signed-rank regression (Bindele & Zhao, 2016). More recently, in the context of missing response under the MNAR assumption, empirical likelihood approaches have been proposed by Niu et al. (2014) and Tang et al. (2014). Their approaches reveal that the considered empirical likelihood functions were defined based on the least-squares estimating equation, which for many of the reasons discussed above is non robust and less efficient in many scenarios.

In this paper, an empirical likelihood approach based on the general rank dispersion function proposed by Jaeckel (1972) is considered in an effort to construct robust confidence regions for the true parameter in model (1.1), where some responses are MNAR. We also investigate the adverse effects of heavy-tailed distributions on the least squares estimator of the regression parameter. The motivation behind the use of the Jaeckel (1972) objective function is threefold: (i) it is known to result in a robust and more efficient estimator compared to many of the mentioned estimation methods such as the least-squares (LS), the maximum likelihood (ML), and many other methods of moments including the least absolute deviation (LAD), (ii) it does not require model

errors' distribution specification, and (iii) similar to the LS approach, it has a simple geometric interpretability. For more on these facts, see Hettmansperger & McKean (2011) and Bindele & Abebe (2012).

The rest of the paper is organized as follows: A weighted empirical likelihood based on the rank-based estimating equation is introduced in Section 2. We also briefly discuss the estimation of the response probability model as proposed by Kim & Yu (2011). In Section 3, we discuss the normal approximation approach as well as the empirical likelihood approach based on imputed residuals. Based on their corresponding influence functions, the robustness of rank-based estimators is discussed in Section 4. To evaluate the performance and the efficiency of the proposed methods, an extensive simulation study is provided in Section 5, and Section 6 gives an illustrative real data example. In Section 7, we provide a conclusion of our findings. In Section 8, assumptions used in the theoretical development, lemmas, and proofs of some obtained results are provided.

2. Weighted empirical likelihood rank based inference

Consider a random sample of size n, $\{(\mathbf{x}_i, y_i),\, i = 1, \cdots, n\}$, from a random vector $(\mathbf{x}, y)$ with distribution $F(\mathbf{x}, y)$, where $\mathbf{x}$ is fully observed but y is subject to missingness. Also, suppose that $\mathbf{x}$ and y are related via the regression model (1.1). Let

$$\delta_i = \begin{cases} 1, & \text{if } y_i \text{ is observed,} \\ 0, & \text{otherwise.} \end{cases}$$

Clearly δ_i is binary and can be assumed Bernoulli distributed with parameter $\pi(\mathbf{x}_i, y_i) = P(\delta_i = 1|\mathbf{x}_i, y_i)$. Also, as in Kim & Yu (2011), δ_i is assumed to be independent of δ_j for all $i \neq j$, and we let $f_j(y_i|\mathbf{x}_i)$ be the conditional distribution of Y_i given $\mathbf{x}_i$ and $\delta_i = j$, for $j = 0, 1$. Note that when $f_0(y_i|\mathbf{x}_i) = f_1(y_i|\mathbf{x}_i)$, we recover the MAR assumption; that is, conditional on $\mathbf{x}_i$, δ_i and y_i are independent.

To construct the weighted empirical likelihood function based on the rank-based estimating equation, and from the inverse marginal probability weighting method introduced by Wang et al. (2004), we consider the random variable $v_i(\boldsymbol{\beta}_0)$ defined by

$$v_i(\boldsymbol{\beta}_0) = \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} \nabla_{\boldsymbol{\beta}} g(\mathbf{x}_i, \boldsymbol{\beta}_0) \varphi\big(R(z_i(\boldsymbol{\beta}_0))/(n+1)\big),$$

where $R(z_i(\boldsymbol{\beta}_0)) = \sum_{j=1}^n I\{z_j(\boldsymbol{\beta}_0) \leq z_i(\boldsymbol{\beta}_0)\}$, $\varphi : (0,1) \to \mathbb{R}$ is a bounded, nondecreasing and square integrable score function, and $z_i(\boldsymbol{\beta}) = y_i - g(\mathbf{x}_i, \boldsymbol{\beta})$. The motivation of considering the above random variable comes from the following rank-based objective function given by

$$D_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} \varphi\big(R(z_i(\boldsymbol{\beta}))/(n+1)\big) z_i(\boldsymbol{\beta}).$$

The rank-based estimator, say, $\widehat{\boldsymbol{\beta}}_n$ is obtained as $\widehat{\boldsymbol{\beta}}_n = \operatorname{Argmin}_{\boldsymbol{\beta} \in \mathscr{B}} D_n(\boldsymbol{\beta})$. The corresponding estimating equation is then given by $n^{-1} \sum_{i=1}^n v_i(\boldsymbol{\beta}) = 0$. Note that $\pi(\mathbf{x}_i, y_i)$ in the expressions above is considered given. When $\pi(\mathbf{x}_i, y_i)$ is unknown with y_i assumed to be missing not at random, the issue is handled and well discussed in Kim & Yu (2011), where the response probability $\pi(\mathbf{x}_i, y_i)$ is assumed to follow a semi-parametric logistic model; that is, $\pi(\mathbf{x}_i, y_i) = \exp\{h(\mathbf{x}_i) + \gamma y_i\}/\big(1 + \exp\{h(\mathbf{x}_i) + \gamma y_i\}\big)$ for some function $h(\cdot)$ and parameter γ. This assumption is reduced to the MAR assumption for $\gamma = 0$. In the same paper, it is demonstrated that $\pi(\mathbf{x}_i, y_i)$ can be consistently estimated by $\widehat{\pi}(\mathbf{x}_i, y_i) = \{1 + \hat{\alpha}(\mathbf{x}_i, \gamma) \exp(-\gamma y_i)\}^{-1}$, where

$$\hat{\alpha}(\mathbf{x}_i, \gamma) = \frac{\sum_{j=1}^n (1 - \delta_i) K_h(\mathbf{x}_i, \mathbf{x}_j)}{\sum_{j=1}^n \delta_i \exp\{-\gamma y_j\} K_h(\mathbf{x}_i, \mathbf{x}_j)},$$

$K_h(\mathbf{t},\mathbf{x}) = h^{-p}K((\mathbf{t}-\mathbf{x})/h^p)$, with $K(\cdot)$ being a kernel function defined on $\mathbb{R}^p$ and $h = h_n$ a bandwith satisfying $h_n \to 0$ and $nh_n^p \to \infty$ as $n \to \infty$. Under assumptions $(I_2) - (I_4)$ given in the Appendix, it is obtained that $\widehat{\pi}(\mathbf{x},y) \to \pi(\mathbf{x},y)$ $a.s.$; see Einmahl & Mason (2005), Rao (2009) and Wied & Weißbach (2012). From assumption (I_5), we have, $\boldsymbol{\beta}_0 = \text{Argmin}_{\boldsymbol{\beta}\in\mathscr{B}} \lim_{n\to\infty} E\{D_n(\boldsymbol{\beta})\}$. This implies that $n^{-1}\sum_{i=1}^n E[v_i(\boldsymbol{\beta}_0)] \to 0$ as $n \to \infty$. To define the empirical likelihood, let $(p_1,\ldots,p_n)^\tau$ denote a vector of probability values satisfying $\sum_i^n p_i = 1$ and $p_i \geq 0$ for all i. Then, the empirical log-likelihood ratio function for $\boldsymbol{\beta}_0$ when γ is assumed known, is given by

$$L(\boldsymbol{\beta},\gamma) = -2\sup\left\{\sum_{i=1}^n \log(np_i) :\ p_i \geq 0, \quad \sum_{i=1}^n p_i = 1 \quad \text{and} \quad \sum_{i=1}^n p_i v_i(\boldsymbol{\beta}_0) = 0\right\},$$

which by the Lagrange multiplier method, gives $p_i = 1/n(1+\boldsymbol{\xi}^\tau v_i(\boldsymbol{\beta}_0))$, with $\boldsymbol{\xi} \in \mathbb{R}^p$ being the Lagrange multiplier parameter. It can also be shown that

$$L(\boldsymbol{\beta}_0,\gamma) = 2\sum_{i=1}^n \log(1+\boldsymbol{\xi}^\tau v_i(\boldsymbol{\beta}_0)). \tag{2.1}$$

The following theorem establishes the asymptotic normality property of $\widehat{\boldsymbol{\beta}}_n$ and the asymptotic distribution of the considered empirical log-likelihood ratio function.

Theorem 1. *Under assumptions $(I_1) - (I_6)$ in the Appendix, we have*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, \gamma_\varphi^{-2}\mathbf{W}_{\boldsymbol{\beta}_0}^{-1}\mathbf{A}_{\boldsymbol{\beta}_0}\mathbf{W}_{\boldsymbol{\beta}_0}^{-1}), \tag{2.2}$$

where $\mathbf{A}_{\boldsymbol{\beta}_0} = E\big[\pi^{-1}(\mathbf{X},Y)\nabla_{\boldsymbol{\beta}}g(\mathbf{X},\boldsymbol{\beta}_0)\nabla_{\boldsymbol{\beta}}^\tau g(\mathbf{X},\boldsymbol{\beta}_0)\big]$ and $\mathbf{W}_{\boldsymbol{\beta}_0} = E[\nabla_{\boldsymbol{\beta}}g(\mathbf{X},\boldsymbol{\beta}_0)\nabla_{\boldsymbol{\beta}}^\tau g(\mathbf{X},\boldsymbol{\beta}_0)]$, and

$$\gamma_\varphi^{-1} = \int_0^1 \varphi(u)\varphi_f(u)du \quad \textit{with} \quad \varphi_f(u) = \frac{f'(F^{-1}(u))}{f(F^{-1}(u))}.$$

Also, if we consider $\widetilde{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta}\in\mathscr{B}}{\text{Argmin}}\, \widetilde{D}_n(\boldsymbol{\beta})$, with $\widetilde{D}_n(\boldsymbol{\beta}) = \dfrac{1}{n}\displaystyle\sum_{i=1}^n \frac{\delta_i}{\widehat{\pi}(\mathbf{x}_i,y_i)}\varphi\begin{pmatrix} R(z_i(\boldsymbol{\beta})) \\ n+1 \end{pmatrix} z_i(\boldsymbol{\beta})$, we would have $\sqrt{n}(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, \gamma_\varphi^{-2}\mathbf{W}_{\boldsymbol{\beta}_0}^{-1}\mathbf{B}_{\boldsymbol{\beta}_0}\mathbf{W}_{\boldsymbol{\beta}_0}^{-1})$, where

$$\begin{aligned}\mathbf{B}_{\boldsymbol{\beta}_0} \quad = \quad & E\left[\pi^{-1}(\mathbf{X},Y)\nabla_{\boldsymbol{\beta}}^\tau g(\mathbf{X},\beta_0)\nabla_{\boldsymbol{\beta}}^\tau g(X,\beta_0)\varphi^2(F(\varepsilon))\right] \\ & + E\{1-\pi^{-1}(X,Y)\nabla_\beta g(X,\beta_0)\nabla_\beta^\tau g(X,\beta_0)E^2[\varphi(F(\varepsilon))|X,\delta=0]\}.\end{aligned}$$

Moreover, for a given γ

$$L(\boldsymbol{\beta}_0,\gamma) \xrightarrow{\mathcal{D}} \chi_p^2 \quad \textit{and} \quad \tilde{L}(\boldsymbol{\beta}_0,\gamma) \xrightarrow{\mathcal{D}} \sum_{i=1}^p \lambda_i\chi_{1,i}^2, \tag{2.3}$$

where the λ_i are the eigenvalues of $\mathbf{B}_{\boldsymbol{\beta}_0}^{1/2}\mathbf{A}_{\boldsymbol{\beta}_0}^{-1}\mathbf{B}_{\boldsymbol{\beta}_0}^{1/2}$, the $\chi_{1,i}^2$ are independent χ^2 distributions with one degree of freedom, and

$$\tilde{L}(\boldsymbol{\beta}_0,\gamma) = 2\sum_{i=1}^n \log(1+\boldsymbol{\xi}^\tau\widetilde{v}_i(\boldsymbol{\beta}_0)) \quad \textit{with} \quad \widetilde{v}_i(\boldsymbol{\beta}_0) = \frac{\delta_i}{\widehat{\pi}(\mathbf{x}_i,y_i)}\nabla_{\boldsymbol{\beta}}g(\mathbf{x}_i,\boldsymbol{\beta}_0)\varphi\big(R(z_i(\boldsymbol{\beta}_0))/(n+1)\big).$$

The proof of this theorem relies on Lemma 1 given in the Appendix. The strong consistency of $\widetilde{\boldsymbol{\beta}}_n$ could be established as in Bindele (2017) with slight modifications.
Based on the empirical log-likelihood, a $(1-\alpha)\times 100\%$ confidence region for $\boldsymbol{\beta}_0$ is given by

$$\mathcal{R}_0 = \left\{\boldsymbol{\beta} : -2\log L(\boldsymbol{\beta},\gamma) \le \chi^2_p(\alpha)\right\} \quad \text{and} \quad \mathcal{R}_1 = \left\{\boldsymbol{\beta} : -2\log \tilde{L}(\boldsymbol{\beta},\gamma) \le \sum_{i=1}^{p}\lambda_i\chi^2_{1,i}(\alpha)\right\},$$

where $\chi^2_p(\alpha)$ is the $(1-\alpha)^{th}$ percentile of the χ^2-distribution with p degrees of freedom, and the $\chi^2_{1,i}(\alpha)$ are the $(1-\alpha)^{th}$ percentiles of the χ^2-distribution with one degree of freedom.

Remark 1. Following Niu et al. (2014), if we were to consider the least squares objective function in Theorem 1 under the model settings, it can be obtained in a straightforward manner that $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{LS}-\boldsymbol{\beta}_0)$ converges in distribution to $N(\mathbf{0},\sigma^{-2}\mathbf{W}^{-1}_{\boldsymbol{\beta}_0}\mathbf{A}_{\boldsymbol{\beta}_0}\mathbf{W}^{-1}_{\boldsymbol{\beta}_0})$, where $\sigma^2 = E(\varepsilon^2|\mathbf{x})$. This results in a relative efficiency of $\sigma^2/\gamma^2_\varphi$ that is shown to be larger than 1 for many of the existing distributions, except for the normal error, where it is about 0.955 (Hettmansperger & McKean, 2011), which demonstrates that the rank-based approach is more efficient than the least squares approach for heavy-tailed model error distributions, and/or for contaminated data. Inference about $\boldsymbol{\beta}_0$ based on equation (2.2) requires a consistent estimator of $\mathbf{W}^{-1}_{\boldsymbol{\beta}_0}\mathbf{A}_{\boldsymbol{\beta}_0}\mathbf{W}^{-1}_{\boldsymbol{\beta}_0}$. Such an estimator can be obtained using sandwich type estimators of $\mathbf{A}_{\boldsymbol{\beta}_0}$ and $\mathbf{W}_{\boldsymbol{\beta}_0}$. That is, based on the expressions of $\mathbf{A}_{\boldsymbol{\beta}_0}$ and $\mathbf{W}_{\boldsymbol{\beta}_0}$, we have $\widehat{\mathbf{A}} = n^{-1}\sum_{i=1}^n \frac{\delta_i}{\pi^2(\mathbf{x}_i,y_i)}\nabla_{\boldsymbol{\beta}} g(\mathbf{x}_i,\widetilde{\boldsymbol{\beta}}_n)\nabla^\tau_{\boldsymbol{\beta}} g(\mathbf{x}_i,\widetilde{\boldsymbol{\beta}}_n)$ and $\widehat{\mathbf{W}} = n^{-1}\sum_{i=1}^n \frac{\delta_i}{\pi(\mathbf{x}_i,y_i)}\nabla_{\boldsymbol{\beta}} g(\mathbf{x}_i,\widetilde{\boldsymbol{\beta}}_n)\nabla^\tau_{\boldsymbol{\beta}} g(\mathbf{x}_i,\widetilde{\boldsymbol{\beta}}_n)$. If we were to consider $\mathbf{W}^{-1}_{\boldsymbol{\beta}_0}\mathbf{B}_{\boldsymbol{\beta}_0}\mathbf{W}^{-1}_{\boldsymbol{\beta}_0}$, similar arguments could be used to estimate $\mathbf{B}_{\boldsymbol{\beta}_0}$.

3. Empirical likelihood based on imputed residuals

Note that although adjusted by the response probability, the vector $v_i(\boldsymbol{\beta}_0)$ is basically defined on observed responses. Thus, its consideration will lead to the complete case analysis, and therefore the information in the data might not be fully explored. To complete the missing responses, we employ two regression imputation methods: the regression simple imputation ($j=1$) and the weighted inverse marginal probability regression imputation ($j=2$) as follows,

$$Z_{ij} = \begin{cases} \delta_i y_i + (1-\delta_i)\, m_0(\mathbf{x}_i), & \text{if } j=1; \\ \dfrac{\delta_i}{\pi(\mathbf{x}_i,y_i)} y_i + \left(1-\dfrac{\delta_i}{\pi(\mathbf{x}_i,y_i)}\right) m_0(\mathbf{x}_i), & \text{if } j=2. \end{cases} \tag{3.1}$$

from which $m_0(\mathbf{x}) = E[Y \mid \mathbf{x}, \delta=0]$ is unknown and needs to be estimated. From a direct application of Bayes' rule, we have for any Borel set B,

$$P(y_i\in B \mid \mathbf{x}_i,\delta_i=0) = P(y_i\in B \mid \mathbf{x}_i,\delta_i=1)\times\frac{(1-\pi(\mathbf{x}_i,y_i))\Delta(\mathbf{x}_i)}{(1-\Delta(\mathbf{x}_i))\pi(\mathbf{x}_i,y_i)},$$

where $\Delta(\mathbf{x}_i) = P(\delta_i=1|\mathbf{x}_i)$. This implies that

$$f_0(y_i \mid \mathbf{x}_i) = f_1(y_i \mid \mathbf{x}_i)\times\frac{O(\mathbf{x}_i,y_i)}{E[O(\mathbf{X}_i,Y_i) \mid \mathbf{x}_i,\delta_i=1]}, \tag{3.2}$$

where $O(\mathbf{x}_i,y_i) = (1-\pi(\mathbf{x}_i,y_i))/\pi(\mathbf{x}_i,y_i)$. Once again, assuming that the response probability model is a semiparametric logistic model, it is obtained in Kim & Yu (2011) that $m_0(\mathbf{x})$ can be estimated by

$$\widehat{m}_0(\mathbf{x};\gamma) = \frac{\sum_{i=1}^n \delta_i y_i K_h(\mathbf{x}_i,\mathbf{x})e^{-\gamma y_i}}{\sum_{i=1}^n \delta_i K_h(\mathbf{x}_i,\mathbf{x})e^{-\gamma y_i}}. \tag{3.3}$$

Theorem 2. *Under assumptions* $(I_2) - (I_4)$ *given in the Appendix, one has* $\widehat{m}_0(\mathbf{x};\gamma) \to m_0(\mathbf{x})$ *a.s., as* $n \to \infty$.

The proof of this result is based on the so-called conditional strong law of large numbers (Rao, 2009). For sake of brevity, it will not be included here. A weak version of Theorem 2 and its corresponding proof can also be found in Kim & Yu (2011). To this end, the imputed responses are obtained as follows

$$\tilde{Z}_{ijn} = \begin{cases} \delta_i y_i + \big(1-\delta_i\big)\widehat{m}_0(\mathbf{x}_i;\gamma) & j=1 \\ \dfrac{\delta_i}{\widehat{\pi}(\mathbf{x}_i,y_i)} y_i + \Big(1 - \dfrac{\delta_i}{\widehat{\pi}(\mathbf{x}_i,y_i)}\Big)\widehat{m}_0(\mathbf{x}_i;\gamma) & j=2. \end{cases} \tag{3.4}$$

Now setting the residuals as $\nu_{ij}(\boldsymbol{\beta}) = \widetilde{Z}_{ijn} - g(\mathbf{x}_i,\boldsymbol{\beta})$, the rank-based objective function is defined as

$$D_n^j(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \varphi\big(R(\nu_{ij}(\boldsymbol{\beta}))/(n+1)\big)\nu_{ij}(\boldsymbol{\beta}),$$

where $R(\nu_{ij}(\boldsymbol{\beta})) = \sum_{k=1}^n I\{\nu_{kj}(\boldsymbol{\beta}) \le \nu_{ij}(\boldsymbol{\beta})\}$ is the rank of $\nu_{ij}(\boldsymbol{\beta})$ among $\nu_{1j}(\boldsymbol{\beta}),\ldots,\nu_{nj}(\boldsymbol{\beta})$, $j=1,2$. The rank-based estimator of $\boldsymbol{\beta}_0$, say $\widehat{\boldsymbol{\beta}}_n^j$ is defined as $\widehat{\boldsymbol{\beta}}_n^j = \mathrm{Argmin}_{\boldsymbol{\beta}\in\mathscr{B}} D_n^j(\boldsymbol{\beta})$.

Remark 2. It is worth pointing out that while γ is assumed to be known in some cases such as in sensitivity analysis or planned missingness, this is not always the case in many other scenarios, and therefore it needs to be estimated. This is a very important issue as γ determines the degree to which the MNAR assumption is satisfied. Based on either an independent survey or a follow-up sample, Kim & Yu (2011) proposed finding γ that solves the following estimating equation

$$\sum_{i=1}^n (1-\delta_i) r_i (y_i - \widehat{m}_0(\mathbf{x}_i;\gamma)) = 0,$$

where $r_i = 1$, if unit i is in the sample and $r_i = 0$, otherwise. Note that for situations where there are outliers in the response space, the above estimating equation will lead to a non robust estimator of γ. We also consider either an independent survey or a follow-up sample, but for robustness purposes, we propose to estimate γ by solving the following estimating equation

$$\sum_{i=1}^n (1-\delta_i)\, r_i \frac{\partial \widehat{m}_0(\mathbf{x}_i;\gamma)}{\partial\gamma} \varphi\big(R(\ell_i(\gamma))/(n_0+1)\big) = 0, \tag{3.5}$$

where $\ell_i(\gamma) = y_i - \widehat{m}_0(\mathbf{x}_i;\gamma)$ and n_0 is the number of nonrespondents. This estimating equation is obtained by taking the negative gradient with respect to γ of $Q(\gamma)$ defined by

$$Q(\gamma) = \sum_{i=1}^n (1-\delta_i)\, r_i \varphi\big(R(\ell_i(\gamma))/(n_0+1)\big)\ell_i(\gamma).$$

Letting $\widehat{\gamma}$ be such an estimator, $\widehat{\gamma} = \mathrm{Argmin}_\gamma\, Q(\gamma)$. Following arguments similar to those in Theorem 1, its asymptotic properties could be obtained in a straightforward manner. As our interest is placed on inference about $\boldsymbol{\beta}_0$, asymptotic properties of $\widehat{\gamma}$ will not be included in this paper. Also, note that if we were to consider $\widehat{\gamma}$ in Theorem 1 and the next theorems, the results might slightly change, as the asymptotic covariance matrices of $\widetilde{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\beta}}_n^j$ could be affected by the asymptotic variance of $\widehat{\gamma}$. For details regarding such a discussion, readers might refer to Kim & Yu (2011) and Niu et al. (2014) who considered the asymptotics under the estimated γ.

3.1. Normal approximation of the rank estimator based on imputed residuals

The normal approximation based inference focuses on the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_n^j$. Denoting $S_n^j(\boldsymbol{\beta})$ to be the negative gradient function of $D_n^j(\boldsymbol{\beta})$, $\widehat{\boldsymbol{\beta}}_n^j$ is a zero of $S_n^j(\boldsymbol{\beta}) = \mathbf{0}$. As in the linear model case (Hettmansperger & McKean, 2011), the distribution of $\widehat{\boldsymbol{\beta}}_n^j$ is strongly related to that of $S_n^j(\boldsymbol{\beta}_0)$. The following theorem establishes the asymptotic distributions of $S_n^j(\boldsymbol{\beta}_0)$ and $\widehat{\boldsymbol{\beta}}_n^j$, respectively.

Theorem 3. *Under assumptions $(I_1) - (I_6)$ in the Appendix,*

$$\sqrt{n} S_n^j(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, \Sigma_{\boldsymbol{\beta}_0}^j) \quad \textit{for } j = 1, 2,$$

where $\mathbf{0}$ is a p-vector of zeros, and $\Sigma_{\boldsymbol{\beta}_0}^j = \lim_{n\to\infty} n^{-1}\boldsymbol{\Sigma}_{jn}\boldsymbol{\Sigma}_{jn}^\tau$, with $\boldsymbol{\Sigma}_{jn}$ defined in Lemma 2 given in the Appendix. Moreover,

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^j - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, \mathbf{M}_j),$$

where $\mathbf{M}_j = \mathbf{V}_j^{-1}\Sigma_{\boldsymbol{\beta}_0}^j\mathbf{V}_j^{-1}$, with $\mathbf{V}_j$ given by

$$\mathbf{V}_j = E\left\{\nabla_{\boldsymbol{\beta}} g(\mathbf{X}, \boldsymbol{\beta}_0)\nabla_{\boldsymbol{\beta}}^\tau g(\mathbf{X}, \boldsymbol{\beta}_0) h^j(\zeta_j(\boldsymbol{\beta}_0))\varphi'\big(H^j(\zeta_j(\boldsymbol{\beta}_0))\big)\right\} + E\{\nabla_{\boldsymbol{\beta}}^2 g(\mathbf{X}, \boldsymbol{\beta}_0)\varphi\big(H^j(\zeta_j(\boldsymbol{\beta}_0))\big)\},$$

$H_i^j(s)$ is the distribution of i.i.d. $\zeta_{ij}(\boldsymbol{\beta}_0) = Z_{ij} - g(\mathbf{x}_i, \boldsymbol{\beta}_0)$, and $h^j(s)$ the corresponding common density for $j = 1, 2$.

It is worth pointing out that the imputation procedure introduces a dependence structure among the residuals. Thus, the residuals given in equation (3.4) are dependent random variables, and therefore, the proof of Theorem 3 above will rely on Lemma 2, which establishes the asymptotic normality property of a statistic defined on dependent random variables.

3.1.1 Estimating the covariance matrix $\mathbf{M}_j$

As discussed earlier, the normal approximation approach is based on the estimated covariance matrix of the rank estimator obtained by minimizing $D_n^j(\boldsymbol{\beta})$. Such a covariance matrix depends on $\Sigma_{\boldsymbol{\beta}_0}^j$ which is a function of the true parameter $\boldsymbol{\beta}_0$, and therefore needs to be estimated. Putting $H_{in}^j(s) = P(\nu_{ij}(\boldsymbol{\beta}_0) \le s)$, it can be shown under the assumptions of Theorem 2 that $H_{in}^j(s) \to H_i^j(s)$ $a.s.$, and by the continuity of φ, we have $\varphi(H_{in}^j(s)) \to \varphi(H_i^j(s))$ $a.s.$ On the other hand, in the proof of Theorem 3, it is shown that $n\boldsymbol{\Sigma}_{jn}^{-1}S_n^j(\boldsymbol{\beta}_0)$ follows a standard multivariate normal distribution, from which we have $\mathrm{Var}\big[\sqrt{n}S_n^j(\boldsymbol{\beta}_0)\big] = n^{-1}\boldsymbol{\Sigma}_{jn}\boldsymbol{\Sigma}_{jn}^\tau$. Also in a matrix form, $S_n^j(\boldsymbol{\beta}_0)$ can be rewritten as $S_n^j(\boldsymbol{\beta}_0) = n^{-1}\nabla_{\boldsymbol{\beta}} g(\mathbf{x}, \boldsymbol{\beta}_0)\boldsymbol{\varphi}(R(\boldsymbol{\nu}_j(\boldsymbol{\beta}_0)))$, where $\boldsymbol{\varphi}(R(\boldsymbol{\nu}_i(\boldsymbol{\beta}_0)))$ is a vector with entries $\varphi(R(\nu_{ij}(\boldsymbol{\beta}_0))/(n+1)$, $i = 1, \ldots, n$. Conditioning on $\mathbf{x}_i$, it is obtained that $\mathrm{Var}\big[\sqrt{n}S_n^j(\boldsymbol{\beta}_0)\big] = n^{-1}\nabla_{\boldsymbol{\beta}} g(\mathbf{x}, \boldsymbol{\beta}_0)\mathrm{Var}\big[\boldsymbol{\varphi}(R(\boldsymbol{\zeta}_j(\boldsymbol{\beta}_0)))\big]\nabla_{\boldsymbol{\beta}}^\tau g(\mathbf{x}, \boldsymbol{\beta}_0)$, for $j = 1, 2$. Thus, the variance of $\sqrt{n}S_n^j(\boldsymbol{\beta}_0)$ differs for $j = 1$ and $j = 2$ only through the distribution inferred by the two different imputation procedures. To this end, as in Brunner & Denker (1994), set $\boldsymbol{\lambda}_i = \nabla_{\boldsymbol{\beta}} g(\mathbf{x}_i, \boldsymbol{\beta}_0)$ and put

$$\begin{aligned}
J_{jn}(s) &= \frac{1}{n}\sum_{i=1}^n H_{in}^j(s), \quad \hat{J}_{jn}(s) = \frac{1}{n}\sum_{i=1}^n I(\nu_{ij}(\boldsymbol{\beta}_0) \le s),\\
F_{jn}(s) &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{\lambda}_i H_{in}^j(s), \quad \hat{F}_{jn}(s) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\lambda}_i I(\nu_{ij}(\boldsymbol{\beta}_0) \le s),\\
\Gamma_n^j(\boldsymbol{\beta}_0) &= S_n^j(\boldsymbol{\beta}_0) - E\big[S_n^j(\boldsymbol{\beta}_0)\big].
\end{aligned}$$

Now, following the same idea in Bindele & Abebe (2015), set

$$\widehat{\mathbf{A}}_{jn} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\lambda}_i \varphi'\Big(\frac{R(\nu_{ij}(\widehat{\boldsymbol{\beta}}_{\varphi}^{j}))}{n+1}\Big) R(\nu_{ij}(\widehat{\boldsymbol{\beta}}_{\varphi}^{j})) \text{ and } \widehat{\boldsymbol{\Sigma}}_{jn} = \widehat{\mathbf{A}}_{jn} - E(\mathbf{A}_j), \tag{3.6}$$

where

$$\begin{aligned} \mathbf{A}_j &= n\int \varphi(J_{jn}(t))\hat{F}_{jn}(dt) + \int \varphi'(J_{jn}(t))\hat{J}_{jn}(t)F_{jn}(dt) \\ &= nS_n^j(\boldsymbol{\beta}_0) + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\lambda}_i \varphi'\Big(\frac{R(\nu_{ij}(\boldsymbol{\beta}_0))}{n+1}\Big) R(\nu_{ij}(\boldsymbol{\beta}_0)). \end{aligned}$$

It is demonstrated in Bindele & Abebe (2015) that $E[\mathbf{A}_j] = n\bar{\boldsymbol{\lambda}}_n[\varphi(1) - \varphi(0)]$, where $\bar{\boldsymbol{\lambda}}_n = n^{-1}\sum_{i=1}^{n}\boldsymbol{\lambda}_i$. This is used in equation (3.6) to approximate $\widehat{\boldsymbol{\Sigma}}_{jn}$ from which the consistency is provided by the following theorem.

Theorem 4. *Letting ς_{jn} be the minimum eigenvalue of $\boldsymbol{\Sigma}_{jn}$ and assuming that $\lim_{n\to\infty} n/\varsigma_{nj} = 0$, we have $\|\widehat{\boldsymbol{\Sigma}}_{jn} - \boldsymbol{\Sigma}_{jn}\| \to 0$ in the L^2-norm as $n \to \infty$. Moreover, from Brunner & Denker (1994), we have $\|n^{-1}\widehat{\boldsymbol{\Sigma}}_{jn}\widehat{\boldsymbol{\Sigma}}_{jn}^{\tau} - \Sigma_{\boldsymbol{\beta}_0}^{j}\| \to 0$ in the L^2-norm as $n \to \infty$.*

The proof of this theorem is a direct consequence of Theorem 3 and is obtained by observing that $\varsigma_{nj} \geq cn^2$ for some positive constant c. On the other hand, $\mathbf{V}_j$ depends on $\boldsymbol{\beta}_0$, and one can estimate $\mathbf{V}_j$ by a sandwich estimator, say $\widehat{\mathbf{V}}_j = \nabla_{\boldsymbol{\beta}} T_n^j(\widehat{\boldsymbol{\beta}}_n^j)$, where $T_n^j(\boldsymbol{\beta}) = \sum_{i=1}^{n}\boldsymbol{\lambda}_i\varphi\big(H_i^j(\zeta_{ij}(\boldsymbol{\beta}))\big)$. From this, the estimated covariance matrix can be defined as $\widehat{\mathbf{M}}_j = \widehat{\mathbf{V}}_j^{-1}\{n^{-1}\widehat{\boldsymbol{\Sigma}}_{jn}\widehat{\boldsymbol{\Sigma}}_{jn}^{\tau}\}\widehat{\mathbf{V}}_j^{-1}$. Combining Theorem 4 with the fact that $\|\widehat{\mathbf{V}}_j - \mathbf{V}_j\| \to 0 \;\; a.s.$, it can be shown that $\widehat{\mathbf{M}}_j \to \mathbf{M}_j$ as $n \to \infty \;\; a.s.$ Hence, a $(1-\alpha)\times 100\%$ normal approximation confidence region for $\boldsymbol{\beta}_0$ with nominal confidence level $1-\alpha$, is given by

$$\mathcal{R}_3^j = \Big\{\boldsymbol{\beta} : (\widehat{\boldsymbol{\beta}}_n^j - \boldsymbol{\beta})\widehat{\mathbf{M}}_j^{-1}(\widehat{\boldsymbol{\beta}}_n^j - \boldsymbol{\beta}) \leq \chi_p^2(\alpha)\Big\}.$$

3.2. Empirical likelihood on imputed residuals

In this section, we adopt the empirical likelihood approach for inference about the true regression parameters. Recall that $S_n^j(\boldsymbol{\beta}) = -\nabla D_n^j(\boldsymbol{\beta})$, which is precisely given by

$$S_n^j(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\varphi\Big(\frac{R(\nu_{ij}(\boldsymbol{\beta}))}{n+1}\Big)\nabla_{\boldsymbol{\beta}} g(\mathbf{x}_i, \boldsymbol{\beta}), \quad j = 1, 2.$$

From this, define the random variable $\eta_{ij}(\boldsymbol{\beta})$ as $\eta_{ij}(\boldsymbol{\beta}) = \varphi\big(R(\nu_{ij}(\boldsymbol{\beta}))/(n+1)\big)\nabla_{\boldsymbol{\beta}} g(\mathbf{x}_i, \boldsymbol{\beta})$ and recall that the rank-based estimator is obtained by solving the estimation equation $S_n^j(\boldsymbol{\beta}) = \mathbf{0}$. Again, under assumption (I_5), $\boldsymbol{\beta}_0 = \text{Argmin}_{\boldsymbol{\beta}\in\mathcal{B}} \lim_{n\to\infty} E\{D_n^j(\boldsymbol{\beta})\}$, which, with probability 1, implies that $E[S_n^j(\boldsymbol{\beta}_0)] \to \mathbf{0}$ as $n \to \infty$. Therefore, the estimating equation $S_n^j(\boldsymbol{\beta}) = \mathbf{0}$ is asymptotically unbiased. Letting $(p_{1j}, \cdots, p_{nj})^{\tau}$ be a vector of probabilities satisfying $\sum_{i=1}^{n} p_{ij} = 1$, with $p_{ij} \geq 0$, $j = 1, 2$, and using the definition of $\eta_{ij}(\boldsymbol{\beta})$, the empirical likelihood ratio at $\boldsymbol{\beta}_0$ is given by

$$R_n^j(\boldsymbol{\beta}_0) = \sup_{(p_{1j},\ldots,p_{nj})\in(0,1)^n} \Big\{\prod_{i=1}^{n}(np_{ij}) : \; \sum_{i=1}^{n} p_{ij} = 1, \; p_{ij} \geq 0, \; \sum_{i=1}^{n} p_{ij}\eta_{ij}(\boldsymbol{\beta}_0) = 0\Big\}. \tag{3.7}$$

From the Lagrange multiplier, it can be shown that $R_n^j(\boldsymbol{\beta}_0)$ is maximized when $p_{ij} = \dfrac{1}{n(1+\boldsymbol{\xi}^\tau \eta_{ij}(\boldsymbol{\beta}_0))}$ with $\boldsymbol{\xi} \in \mathbb{R}^d$ satisfying the following nonlinear equation:

$$h(\boldsymbol{\xi}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\eta_{ij}(\boldsymbol{\beta}_0)}{1+\boldsymbol{\xi}^\tau \eta_{ij}(\boldsymbol{\beta}_0)} = \mathbf{0}. \tag{3.8}$$

Now, combining equations (3.7) and (3.8) gives,

$$-2\log R_n^j(\boldsymbol{\beta}_0) = -2\log \prod_{i=1}^{n} \left(1+\boldsymbol{\xi}^\tau \eta_{ij}(\boldsymbol{\beta}_0)\right)^{-1} = 2\sum_{i=1}^{n} \log\left(1+\boldsymbol{\xi}^\tau \eta_{ij}(\boldsymbol{\beta}_0)\right), \tag{3.9}$$

which leads to the following theorem

Theorem 5. *Under assumptions $(I_1)-(I_6)$ in the Appendix, one has*

$$-2\log R_n^j(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} \chi_p^2 \quad as \quad n\to\infty.$$

The empirical likelihood (EL) confidence region for $\boldsymbol{\beta}_0$ with nominal confidence level $1-\alpha$, is given by

$$\mathcal{R}_4^j = \{\boldsymbol{\beta} : -2\log R_n^j(\boldsymbol{\beta}) \le \chi_p^2(\alpha)\}.$$

This confidence region enables us to perform statistical inference about $\boldsymbol{\beta}_0$.

4. Robustness

To access the robustness of the rank-based approach, we derive the influence functions that results from the considered objective functions. From Theorems 1 and 3, it is not difficulty to see that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = (\gamma_\varphi \mathbf{W}_{\boldsymbol{\beta}_0})^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} \nabla_{\boldsymbol{\beta}} g(\mathbf{x}_i, \boldsymbol{\beta}_0) \varphi\left(\frac{R(z_i(\boldsymbol{\beta}_0))}{n+1}\right) + o_p(1),$$

and similarly

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^j - \boldsymbol{\beta}_0) = \mathbf{V}_j^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla_{\boldsymbol{\beta}} g(\mathbf{x}_i, \boldsymbol{\beta}_0) \varphi\left(\frac{R(\nu_{ij}(\boldsymbol{\beta}_0))}{n+1}\right) + o_p(1).$$

Following Bindele & Abebe (2012), the influence functions of both $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\beta}}_n^j$ are obtained as

$$\mathrm{IF}(\mathbf{x}, y) = \frac{\delta(\gamma_\varphi \mathbf{W}_{\boldsymbol{\beta}_0})^{-1}}{\pi(\mathbf{x}, y)} \nabla_{\boldsymbol{\beta}} g(\mathbf{x}, \boldsymbol{\beta}_0) \varphi\left(F(\varepsilon)\right) \quad \text{and} \quad \mathrm{IF}_j(\mathbf{x}, y) = \mathbf{V}_j^{-1} \nabla_{\boldsymbol{\beta}} g(\mathbf{x}, \boldsymbol{\beta}_0) \varphi(H^j(\zeta_j(\boldsymbol{\beta}_0))),$$

respectively. From assumptions (I_1), (I_2), (I_4) and (I_6) in the Appendix, it can be shown in a straightforward manner that $\mathrm{IF}(\mathbf{x}, y)$ and $\mathrm{IF}_j(\mathbf{x}, y)$ are bounded in the y-space, and almost surely bounded in the $\mathbf{x}$-space. Thus, the corresponding estimators are robust to outlying observations in the response space.

Remark 3. From the LS objective, just considering the weighted version as discussed in Remark 1, one would obtain

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}_0) = (\sigma \mathbf{W}_{\boldsymbol{\beta}_0})^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} \nabla_{\boldsymbol{\beta}} g(\mathbf{x}_i, \boldsymbol{\beta}_0)[y_i - g(\mathbf{x}_i, \boldsymbol{\beta}_0)] + o_p(1),$$

from which following similar arguments as in Bindele & Abebe (2012), results in

$$\mathrm{IF}_{LS}(\mathbf{x}, y) = (\sigma \mathbf{W}_{\boldsymbol{\beta}_0})^{-1} \frac{\delta}{\pi(\mathbf{x}, y)} \nabla_{\boldsymbol{\beta}} g(\mathbf{x}, \boldsymbol{\beta}_0)[y - g(\mathbf{x}, \boldsymbol{\beta}_0)].$$

Under the assumptions, it is not hard to see that while this influence function is almost surely bounded in the $\mathbf{x}$-space, it is unbounded in the y-space. Thus, $\widehat{\boldsymbol{\beta}}_{LS}$ is not robust to outlying observations in the y-space.

5. Simulation Study

5.1. Simulation settings

In order to confirm the validity of our theoretical findings and to show the performance of the empirical likelihood rank-based approach compared to the normal approximation approach, an extensive simulation under different settings is conducted from which coverage probabilities (CP) and average lengths (AL) of confidence intervals/regions of the true regression coefficients are calculated. In model (1.1), we consider the simple regression function g defined as $g(x, \boldsymbol{\beta}) = \beta_1 + \beta_2 x$ with $\boldsymbol{\beta} = (\beta_1, \beta_2) = (1.7, 0.7)$. The random errors ε are generated from the contaminated normal distribution $\mathcal{CN}(\epsilon, \sigma) = \epsilon N(0, 1) + (1 - \epsilon) N(1, \sigma^2)$ for different rates of contamination ($\epsilon = 0, 0.3, 0.5$) with $\sigma = 2$, the t-distribution with various degrees of freedom ($df = 5, 15, 25, 40, 50$) with sample size $n = 200$, and the Laplace distributions with different sample sizes ($n = 15, 50, 100, 250$). These distributions are chosen to be able to study the effect of contamination and tail thickness, respectively. This same choice covers the standard normal distribution in $\mathcal{CN}(\epsilon)$ by setting $\epsilon = 0$ or under the t-distribution when the degree of freedom gets larger and larger. The Laplace distribution on the other hand, allows us to study the effect of the sample size on coverage probabilities and average lengths of the confidence intervals of β_2. The covariate x is generated from $N(1, 1)$ and δ is generated from a Bernoulli distribution with response probability $\pi(x, y)$. To acomodate a nonlinear case, as in Bindele & Zhao (2016), we also consider the Micheaelis-Menten function defined as $g(x, \beta) = x/(\beta + x)$, where the true $\beta = 1$, x is generated from an exponential distribution, and the random errors are generated just from $\mathcal{CN}(0.9)$ and t_3 for the sake of brevity. We investigate five different response probability cases:

Case 1: $\pi(x, y) = 1/\big(1 + \exp\{0.35 - x - 0.8y - 0.1y^2\}\big)$.

Case 2: $\pi(x, y) = 1/\big(1 + \exp\{0.15 - 0.1x - 0.6y + 0.9xy\}\big)$.

Case 3: $\pi(x, y) = 1/\big(1 + \exp\{-0.3\exp(x) - 0.1y\}\big)$.

Case 4: $\pi(x, y) = \exp\{-0.5x + 0.4x^2 + 0.3y\}/\big(1 + \exp\{-0.5x + 0.4x^2 + 0.3y\}\big)$.

Case 5: $\pi(x, y) = \exp\{-0.8\sin x + 0.6y\}/\big(1 + \exp\{-0.8\sin x + 0.6y\}\big)$.

While Cases 3–5 satisfy the assumed response probability assumption with γ set at 0.1, 0.3, and 0.6, respectively, Cases 1–2, which do not satisfy such an assumption, are used to examine the robustness of the proposed estimator against departure from the assumed missing assumption. Cases 1, 2, 4, and 5 give on average a response probability of roughly about 70%, while Case 3 gives on average a response probability of about 60%. As in real data situations, γ is unspecified, we estimate γ by solving equation (3.5) via either the Newton-Raphson or the Bisection approaches, where the follow-up rate used is set at 30%. The corresponding estimates ($\widehat{\gamma}$) are 0.098, 0.307, and 0.598. The choice of the kernel function having less importance (Einmahl & Mason, 2005), we consider the Epanechnikov kernel function, that is, $K(u) = 0.75(1 - u^2)I(|u| \leq 1)$. On the other hand, as the estimation of both $\pi(\mathbf{x}, y)$ and $m_0(\mathbf{x}, \gamma)$ involve selecting the bandwidth, similar to Delecroix et al. (2006), we consider a joint minimization of $D_n^j(\boldsymbol{\beta}, h)$, where the starting value of

$h \in \{h : c_1 n^{-\alpha_1} < h < c_2 n^{-\alpha_2}\}$, for some $c_1, c_2 > 0$, $1/8 < \alpha_1 < \alpha_2 < 1/4$. The score function φ that appears in the theoretical development of the paper is taken to be the Wilcoxon score function; that is $\varphi(u) = \sqrt{12}(u - 1/2)$.

From 5000 replications, coverage probabilities (CP) and average lengths (AL) of the true slope β_2 based on the EL approach are reported and are compared with those based on the normal approximation (NA) approach. The approaches considered are: the least squares (LS) based on the normal approximation under regression simple imputation (SI-NA$_{LS}$) and the weighted inverse marginal probability regression imputation (IP-NA$_{LS}$), the rank-based approach with respect to the normal approximation under regression simple imputation (SI-NA$_R$) and the weighted inverse marginal probability regression imputation (IP-NA$_R$), the empirical likelihood based on the LS estimating equation under regression simple imputation (SI-EL$_{LS}$) and the weighted inverse marginal probability regression imputation (IP-EL$_{LS}$), and the empirical likelihood based on the rank estimating equation under regression simple imputation (SI-EL$_R$) and the weighted inverse marginal probability regression imputation (IP-EL$_R$). Also, the weighted rank-based normal approximation (WNA$_R$) and the weighted empirical likelihood based on the weighted rank-based estimating equation (WEL$_R$) using $\mathcal{R}_1$ were considered. The CP and AL based on the EL approach for both the LS and the R are obtained with respect to their corresponding objective functions, while those based on the NA approach are based on the estimated covariance matrices from the LS and the R estimators. The results of the simulation study are displayed in Tables 1 – 8.

Insert Tables 1-8 about here.

5.2. Discussion

From Tables 1 and 2, it is observed that based on either the regression simple imputation (SI) or the weighted inverse marginal probability regression imputation (IP), while the EL approach based on LS estimating equation (SI-EL$_{LS}$, IP-EL$_{LS}$) provides better coverage probabilities compared to its normal approximation (SI-NA$_{LS}$, IP-NA$_{LS}$) counterpart, it has a similar performance with the rank-based normal approximation approach (SI-NA$_R$, IP-NA$_R$). The same observation is made when it comes to average lengths of confidence intervals, but with the EL based on the LS providing slightly shorter average lengths compared to the rank-based normal approximation. These methods give larger coverage probabilities for small degrees of freedom, and such coverage probabilities converge to the nominal confidence level as the degrees of freedom increase. On the other hand, the EL rank-based approach (SI-EL$_R$, IP-EL$_R$, WEL$_R$) gives consistent coverage probabilities that are closer to the nominal confidence level compared to its NA (SI-NA$_R$, IP-NA$_R$, WNA$_R$) competitor. One also observes that except for the weighted rank-based EL (WEL$_R$) approach, the SI-EL$_R$ and IP-EL$_R$ remain superior to all, both in terms of coverage probabilities and average lengths of confidence intervals.

Considering the contaminated normal distribution model error (Tables 3 and 4), based either on SI or IP, once again, the EL approach based on LS estimating equation (SI-EL$_{LS}$, IP-EL$_{LS}$) provides better coverage probabilities compared to its normal approximation (SI-NA$_{LS}$, IP-NA$_{LS}$) counterpart. Although slightly better, such a performance is comparable to that of the rank-based normal approximation approach (SI-NA$_R$, IP-NA$_R$). At $\epsilon = 0$, all the considered methods provide coverage probabilities close to the nominal confidence level and smaller average lengths. This is not surprising as $\epsilon = 0$ corresponds to the standard normal distribution model error. Also, while the rank-based empirical likelihood methods (SI-EL$_R$, IP-EL$_R$, WEL$_R$) show their superiority by giving consistent coverage probabilities close to the nominal confidence level and shorter average lengths, the other approaches provide larger average lengths and coverage probabilities that are increasing as the rate of contamination increases. It is worth pointing out that, generally, average lengths increase as the rate of contamination increases.

When it comes to the Laplace distribution model error, under the five cases and based either on SI or IP as can be seen in Tables 5 and 6, similar observations are made as in the previous two distributions model error.

As the effect of the sample size is of interest in this scenario, for all the considered approaches, we observe that as the sample size increases, the coverage probabilities converge to the nominal confidence level and the average lengths decrease, as expected, with the rank-based empirical likelihood showing its dominance over all the other approaches.

Similar observations are made for the nonlinear Micheaelis-Menten model under the considered model error distributions, as can be seen in Tables 7 and 8.

Generally, the contaminated normal distribution provides shorter average lengths compared to the other distributions considered, and the IP provides shorter average lengths compared to the SI. Also, average lengths obtained based on the weighted EL are slightly shorter compared to those obtained via the imputed EL. As the imputed approaches explore the entire data, and their performance being very similar to the weighted empirical likelihood, in practice, it would be preferable to use the imputed rank-based empirical likelihood. Finally, it is observed that Case 3 results in larger average lengths and coverage probabilities compared to the other four cases. One suspects that this might be caused by a larger percentage of nonrespondents in Case 3.

6. Real Data Example

Consider data from the statistical consulting center project performed at the Department of Mathematics and Statistics of the University of South Alabama. This data came from Cobb County, GA , Women, Infants, and Children (WIC) program and is used here with permission of the investigators. The data contain about 2500 observations on six variables: neonatal baby weight (y), age (x_1), body mass index (BMI, x_2), smoking status (x_3), and indicators for race (x_4) and Hispanic ethnicity (x_5). The purpose of the study is to investigate how accurately neonatal baby weight can be predicted based on body mass index, smoking status (yes or no), race (white or black) and Hispanic (yes or no) of the mother by fitting a linear model. The motivation of using our proposed approach comes from the fact that the response of interest (neonatal baby weight) contains approximately 43% of missing data. Also, one might except that mothers with premature babies would be less likely to disclose their baby's weight. Thus, the missingness in such a case depends on the response itself, and therefore, motivating the use of the proposed approach. From the nonrespondents, 25% were randomly selected for follow-up samples. This represents about 1450 respondents from the original data, and about 269 who responded to the follow-up. The parameter γ was estimated following equation (3.5) using the 269 observations from the follow-up sample. The outputs of the analysis are displayed in Table 9 and Figure 1 below.

Insert Table 9 and Figure 1 about here.

From the studentized residuals plots and the residuals Q-Q plots (Figure 1) of the LS and rank-based (R) on the complete case (CC) analysis, it seems that there exist a few outliers, and also the model error could be approximated by a normal distribution, which would make the LS more appropriate for the NA approach. This is confirmed from the output in Table 9, as in the CC case, the NA_{LS} performs slightly better than the NA_R, but the EL_R outperforms the EL_{LS}. However, the estimates obtained from the CC analysis may be biased because the analysis does not take into account the missing information (Little & Rubin, 2002). The bias is reduced when the adjustment is made considering the response probability, as it can be seen from the weighted complete case (WCC) analysis. This shows that with such an adjustment the rank-based approach outperforms the LS in terms of average lengths of the confidence intervals using either the NA approach or the EL approach, with a better performance by the EL approach. When the missing responses are imputed using either the SI or the IP, the EL based on both the LS and rank-based provides smaller average lengths of confidence intervals compared to their normal approximation counterpart, with a better performance for the EL based on the rank-based estimating equation. Also the EL based on IP provided smaller average lengths

compared to that based on SI, as noticed in the simulation study.

7. Conclusion

Overall, it is not surprising that for heavy-tailed model errors and contaminated data such as in the presence of outliers, the rank-based approach provides robust and more efficient estimators than its least-squares counterpart (Hettmansperger & McKean, 2011; Bindele & Abebe, 2012) for the complete case analysis. However, when it comes to direct statistical inference (confidence intervals/regions) about the true regression coefficients from model (1.1) with responses missing not at random, via the simulation study and the real data example in this paper, it is proven that the empirical likelihood based on the rank-based estimating equation provides a more appealing alternative compared to its normal approximation inference and its least squares counterpart. As the approaches considered utilize estimating equations (EE), where an EE approach is a natural solution for longitudinal data. For future work, it is of interest to generalize these methods to longitudinal data models with MNAR.

Acknowledgement

The authors are grateful to the anonymous reviewers, the Associate Editor and the Editor for their valuable comments and suggestions that helped to improve the paper. Yichuan Zhao was supported by the NSA grant.

8. Appendix

This Appendix provides assumptions used in the development of theoretical results, lemmas and proofs of some of the results.

8.1. Assumptions

(I_1) φ is a nondecreasing, bounded and twice continuously differentiable score function defined on $(0,1)$ that can be standardized as:

$$\int_0^1 \varphi(u)du = 0 \quad and \quad \int_0^1 \varphi^2(u)du = 1,$$

and the model error conditional on the covariates has a distribution with a finite Fisher information.

(I_2) $g(\cdot)$ being a function of two variables $\mathbf{x}$ and $\boldsymbol{\beta}$, it is required that g has continuous derivatives with respect to $\boldsymbol{\beta}$ that are bounded up to order 3 by p-integrable functions of $\mathbf{x}$, independent of $\boldsymbol{\beta}$, $p \geq 1$.

(I_3) $K(\cdot)$ is of bounded variation smooth kernel function with bandwidth h_n satisfying $nh_n^{4r} \to 0$, where r is the order of smoothness of $K(\cdot)$. Also, there exists $c > 0$ such that $c(\log n/n)^{1-2/p} < h_n$, with $p > 2$ and $h_n \to 0$ as $n \to \infty$.

(I_4) $\sup_{\mathbf{x}} E[|Y|^p|\mathbf{X} = \mathbf{x}] < \infty$, for $p \geq 1$. There exists a positive constant c such that $\pi(\mathbf{x}, y) \geq c > 0$ and $E(\pi(\mathbf{X}, Y)|\mathbf{X}) \neq 1$, for all $\mathbf{x}, y$. Also, $E(\exp\{2\gamma Y\}) < \infty$ and $E(\exp\{\lambda X\}) < \infty$, for some λ.

(I_5) For fixed n, $\boldsymbol{\beta}_{0,n}, \boldsymbol{\beta}_{0,n}^j \in Int(\mathcal{B})$ are the unique minimizers of $E[D_n(\boldsymbol{\beta})]$ and $E[D_n^j(\boldsymbol{\beta})]$, respectively such that $\lim_{n\to\infty} \boldsymbol{\beta}_{0,n} = \boldsymbol{\beta}_0$ and $\lim_{n\to\infty} \boldsymbol{\beta}_{0,n}^j = \boldsymbol{\beta}_0$, for $j = 1, 2$.

(I_6) $\mathbf{A}_{\boldsymbol{\beta}_0}$, $\mathbf{B}_{\boldsymbol{\beta}_0}$, $\mathbf{W}_{\boldsymbol{\beta}_0}$, $\Sigma_{\boldsymbol{\beta}_0}^j$, and $\mathbf{V}_j$, for $j = 1, 2$, are assumed to be positive definite.

Assumption (I_1) is a regular assumption in the rank-based framework; see Hettmansperger & McKean (2011) and Bindele & Abebe (2012). Assumptions $(I_2)-(I_4)$ are necessary to ensure the result in Theorem 2; see Einmahl & Mason (2005), Rao (2009) and Wied & Weißbach (2012). The identifiability condition (I_5) is a key to ensure the strong consistency of the rank-based estimator; see Bindele (2017). Finally, assumptions (I_6) together with the previous assumptions are needed to establish the $\sqrt{n}-$ asymptotic normality of the proposed estimators.

Lemma 1. *Under assumptions $(I_1)-(I_6)$ in the Appendix, we have*

(i) $n^{-1}\sum_{i=1}^n(\widetilde{v}_i(\boldsymbol{\beta}_0)-v_i(\boldsymbol{\beta}_0))\to 0 \quad a.s.$

(ii) $n^{-1/2}\sum_{i=1}^n v_i(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0},\mathbf{A}_{\boldsymbol{\beta}_0})$ *and* $n^{-1}\sum_{i=1}^n v_i(\boldsymbol{\beta}_0)v_i^\tau(\boldsymbol{\beta}_0)\xrightarrow{P}\mathbf{A}_{\boldsymbol{\beta}_0}$.

(iii) $n^{-1/2}\sum_{i=1}^n \widetilde{v}_i(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0},\mathbf{B}_{\boldsymbol{\beta}_0})$ *and* $n^{-1}\sum_{i=1}^n \widetilde{v}_i(\boldsymbol{\beta}_0)\widetilde{v}_i^\tau(\boldsymbol{\beta}_0)\xrightarrow{P}\mathbf{A}_{\boldsymbol{\beta}_0}$.

The following lemma establishes the asymptotic normality of a statistic defined on dependent random variables.

Lemma 2 (Brunner & Denker (1994))**.** *Let ς_{jn} be the minimum eigenvalue of $\boldsymbol{\Sigma}_{jn}=Var(U_{jn})$ with U_{jn} given by*

$$U_{jn}=\int\varphi(J_{jn}(s))(\hat{F}_{jn}-F_{jn})(ds)+\int\varphi'(J_{jn}(s))(\hat{J}_{jn}(s)-J_{jn}(s))F_{jn}(ds).$$

Suppose that $\varsigma_{jn}\geq cn^a$ for some constants $c,a\in\mathbb{R}$ and $m(n)$ is such that $M_0n^\alpha\leq m(n)\leq M_1n^\alpha$ for some constants $0<M_0\leq M_1<\infty$ and $0<\alpha<(a+1)/2$. Then $m(n)\boldsymbol{\Sigma}_{jn}^{-1}\Gamma_n^j(\boldsymbol{\beta}_0)$ is asymptotically standard multivariate normal, provided φ is twice continuously differentiable with bounded second derivative.

The proof of this lemma can be constructed along the lines of that of Theorem 3.1 in Brunner & Denker (1994). The proof will not be included here, and readers seeking for details are referred to the aforementioned paper.

8.2. Proofs

Proof of Lemma 1. Set $\widetilde{S}_n(\boldsymbol{\beta}_0)=n^{-1}\sum_{i=1}^n\widetilde{v}_i(\boldsymbol{\beta}_0)$ and $S_n(\boldsymbol{\beta}_0)=n^{-1}\sum_{i=1}^n v_i(\boldsymbol{\beta}_0)$. Put $a_{ni}=\dfrac{R(z_i(\boldsymbol{\beta}_0))}{n+1}$.

(i) We have $\widetilde{S}_n(\boldsymbol{\beta}_0)-S_n(\boldsymbol{\beta}_0)=\dfrac{1}{n}\sum_{i=1}^n\delta_i\nabla_{\boldsymbol{\beta}}g(\mathbf{x}_i,\boldsymbol{\beta}_0)\varphi(a_{ni})\left[\dfrac{1}{\widehat{\pi}(\mathbf{x}_i,y_i)}-\dfrac{1}{\pi(\mathbf{x}_i,y_i)}\right]$, and

$$\|\widetilde{S}_n(\boldsymbol{\beta}_0)-S_n(\boldsymbol{\beta}_0)\|\leq\frac{1}{n}\sum_{i=1}^n\delta_i\|\nabla_{\boldsymbol{\beta}}g(\mathbf{x}_{i,\boldsymbol{\beta}_0})\||\varphi(a_{ni})|\left|\frac{1}{\widehat{\pi}(\mathbf{x}_i,y_i)}-\frac{1}{\pi(\mathbf{x}_i,y_i)}\right|.$$

From the boundedness of φ, there exists a positive constant c_0 such that $|\varphi(t)|\leq c_0$ for all $t\in(0,1)$. Thus,

$$\|\widetilde{S}_n(\boldsymbol{\beta}_0)-S_n(\boldsymbol{\beta}_0)\|\leq\frac{c_0}{n}\sum_{i=1}^n\|\nabla_{\boldsymbol{\beta}}g(\mathbf{x}_i,\boldsymbol{\beta}_0)\|\left|\frac{1}{\widehat{\pi}(\mathbf{x}_i,y_i)}-\frac{1}{\pi(\mathbf{x}_i,y_i)}\right|.$$

Applying Cauchy-Schwarz inequality to the right hand side of this inequality gives

$$\begin{aligned}\|\widetilde{S}_n(\boldsymbol{\beta}_0) - S_n(\boldsymbol{\beta}_0)\| &\leq c_0\left[\frac{1}{n}\sum_{i=1}^n \|\nabla_{\boldsymbol{\beta}} g(\mathbf{x}_i,\boldsymbol{\beta}_0)\|^2\right]^{1/2}\left[\frac{1}{n}\sum_{i=1}^n\left|\frac{1}{\widehat{\pi}(\mathbf{x}_i,y_i)} - \frac{1}{\pi(\mathbf{x}_i,y_i)}\right|^2\right]^{1/2}\\ &\leq c_0\left[\frac{1}{n}\sum_{i=1}^n \|\nabla_{\boldsymbol{\beta}} g(\mathbf{x}_i,\boldsymbol{\beta}_0)\|^2\right]^{1/2}\left[\max_{1\leq i\leq n}\left|\frac{1}{\widehat{\pi}(\mathbf{x}_i,y_i)} - \frac{1}{\pi(\mathbf{x}_i,y_i)}\right|^2\right]^{1/2}.\end{aligned}$$

By the strong law of large numbers, $n^{-1}\sum_{i=1}^n \|\nabla_{\boldsymbol{\beta}} g(\mathbf{x}_i,\boldsymbol{\beta}_0)\|^2 \xrightarrow{a.s.} E[\|\nabla_{\boldsymbol{\beta}} g(\mathbf{X}_i,\boldsymbol{\beta}_0)\|^2] < \infty$ by assumption (I_2). On the other hand, from the fact $\widehat{\pi}(\mathbf{x}_i,y_i) \to \pi(\mathbf{x}_i,y_i)$ $a.s.$, for each i under assumptions $(I_2)-(I_4)$ (Einmahl & Mason, 2005; Rao, 2009; Wied & Weißbach, 2012), we have

$$\max_{1\leq i\leq n}\left|\frac{1}{\widehat{\pi}(\mathbf{x}_i,y_i)} - \frac{1}{\pi(\mathbf{x}_i,y_i)}\right|^2 \xrightarrow{a.s.} 0. \tag{8.1}$$

To see this, note that the fact that $\widehat{\pi}(\mathbf{x}_i,y_i) \to \pi(\mathbf{x}_i,y_i)$ $a.s.$ and $\pi(\mathbf{x}_i,y_i) \geq c > 0$ for all i, by assumptions $(I_2)-(I_4)$, implies that with probability 1, given $\epsilon^\star = c/2 > 0$, there exists an integer $N(\epsilon^\star) > 0$ such that $|\widehat{\pi}(\mathbf{x}_i,y_i) - \pi(\mathbf{x}_i,y_i)| < \epsilon^\star$, for all $n \geq N(\epsilon^\star)$. Moreover, with probability 1, we have $||\widehat{\pi}(\mathbf{x}_i,y_i)| - |\pi(\mathbf{x}_i,y_i)|| \leq |\widehat{\pi}(\mathbf{x}_i,y_i) - \pi(\mathbf{x}_i,y_i)| < \epsilon^\star$ so that $|\widehat{\pi}(\mathbf{x}_i,y_i)| > c - \epsilon^\star = c/2$, for all $n \geq N(\epsilon^\star)$ and for each i. Now, let $\epsilon > 0$ be arbitrary. Then, with probability 1, there exists an integer $N(\epsilon)$ such that for all $n \geq N(\epsilon)$, $|\widehat{\pi}(\mathbf{x}_i,y_i) - \pi(\mathbf{x}_i,y_i)| < \epsilon c^2/2$, for all i. Setting $N = \max\{N(\epsilon^\star), N(\epsilon)\}$, with probability 1, we have for all $n \geq N$,

$$\left|\frac{1}{\widehat{\pi}(\mathbf{x}_i,y_i)} - \frac{1}{\pi(\mathbf{x}_i,y_i)}\right| = \frac{|\widehat{\pi}(\mathbf{x}_i,y_i) - \pi(\mathbf{x}_i,y_i)|}{|\widehat{\pi}(\mathbf{x}_i,y_i)|\pi(\mathbf{x}_i,y_i)} = |\widehat{\pi}(\mathbf{x}_i,y_i) - \pi(\mathbf{x}_i,y_i)|\frac{1}{|\widehat{\pi}(\mathbf{x}_i,y_i)|}\frac{1}{\pi(\mathbf{x}_i,y_i)} < \frac{\epsilon}{2}c^2\cdot\frac{2}{c^2} = \epsilon,$$

for all i. Thus, $\max_{1\leq i\leq n}\left|\frac{1}{\widehat{\pi}(\mathbf{x}_i,y_i)} - \frac{1}{\pi(\mathbf{x}_i,y_i)}\right| \to 0$ $a.s.$, and therefore, $\widetilde{S}_n(\boldsymbol{\beta}_0) - S_n(\boldsymbol{\beta}_0) \xrightarrow{a.s.} 0$.

(ii) Note that $E\big[\varphi(R(z_i(\boldsymbol{\beta}_0))/(n+1))\big] = n^{-1}\sum_{i=1}^n \varphi(i/(n+1)) \to \int_0^1 \varphi(t)dt = 0$ as $n \to \infty$, by assumption (I_1). This together with the fact that $\boldsymbol{\beta}_0 = \text{Argmin}_{\boldsymbol{\beta}\in\mathscr{B}} \lim_{n\to\infty} E[D_n(\boldsymbol{\beta})]$ implies that $E[S_n(\boldsymbol{\beta}_0)] \to 0$ as $n \to \infty$. On the other hand, it can be shown under assumption (I_1) that $\text{Var}\big[\varphi(R(z_i(\boldsymbol{\beta}_0))/(n+1))\big] = n^{-1}\sum_{i=1}^n \varphi^2(i/(n+1)) \to 1$ as $n \to \infty$. Also, following Hettmansperger & McKean (2011),

$$\begin{aligned}\text{cov}\left[\varphi\Big(\frac{R(z_i(\boldsymbol{\beta}_0))}{n+1}\Big), \varphi\Big(\frac{R(z_j(\boldsymbol{\beta}_0))}{n+1}\Big)\right] &= \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{\substack{j=1\\ i\neq j}}^n \varphi\Big(\frac{i}{n+1}\Big)\varphi\Big(\frac{j}{n+1}\Big)\\ &= -\frac{1}{n(n-1)}\sum_{i=1}^n \varphi^2\Big(\frac{i}{n+1}\Big) \to 0 \text{ as } n\to\infty\end{aligned}$$

as $n^{-1}\sum_{i=1}^n \varphi^2\big(i/(n+1)\big) \to \int_0^1 \varphi^2(t)dt = 1$, by assumption (I_1). Thus, conditional on $\mathbf{x}_i$,

$$\begin{aligned}\text{Var}\big[\sqrt{n}S_n(\boldsymbol{\beta}_0)\big] &= \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi^2(\mathbf{x}_i,y_i)}\nabla_{\boldsymbol{\beta}} g(\mathbf{x}_i,\boldsymbol{\beta}_0)\nabla_{\boldsymbol{\beta}}^\tau g(\mathbf{x}_i,\boldsymbol{\beta}_0)\text{Var}\big[\varphi(R(z_i(\boldsymbol{\beta}_0))/(n+1))\big]\\ &\quad + \frac{2}{\pi(\mathbf{x}_i,y_i)\pi(\mathbf{x}_j,y_j)}\text{cov}\left[\varphi\Big(\frac{R(z_i(\boldsymbol{\beta}_0))}{n+1}\Big), \varphi\Big(\frac{R(z_j(\boldsymbol{\beta}_0))}{n+1}\Big)\right]\\ &= \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi^2(\mathbf{x}_i,y_i)}\nabla_{\boldsymbol{\beta}} g(\mathbf{x}_i,\boldsymbol{\beta}_0)\nabla_{\boldsymbol{\beta}}^\tau g(\mathbf{x}_i,\boldsymbol{\beta}_0) + o(1) \quad \text{with probability 1.}\end{aligned}$$

Thus, $\mathrm{Var}\big[\sqrt{n}S_n(\boldsymbol{\beta}_0)\big] \xrightarrow{a.s.} \mathbf{A}_{\boldsymbol{\beta}_0} = E\big[\pi^{-1}(\mathbf{X},Y)\nabla_{\boldsymbol{\beta}}g(\mathbf{X},\boldsymbol{\beta}_0)\nabla_{\boldsymbol{\beta}}^{\tau}g(\mathbf{X},\boldsymbol{\beta}_0)\big]$. Now, putting

$$T_n(\boldsymbol{\beta}_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(\mathbf{x}_i,y_i)}\nabla_{\boldsymbol{\beta}}g(\mathbf{x}_i,\boldsymbol{\beta}_0)\varphi(F(z_i(\boldsymbol{\beta}_0))), \tag{8.2}$$

and as in the proof of Theorem 3.5.2 in Hettmansperger & McKean (2011), one can obtain that $\sqrt{n}\big[S_n(\boldsymbol{\beta}_0) - T_n(\boldsymbol{\beta}_0)\big] \xrightarrow{P} 0$. From a direct application of the central limit theorem, we have $\sqrt{n}T_n(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0},\mathbf{A}_{\boldsymbol{\beta}_0})$. Thus, $\sqrt{n}S_n(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0},\mathbf{A}_{\boldsymbol{\beta}_0})$. On the other hand,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}v_i(\boldsymbol{\beta}_0)v_i^{\tau}(\boldsymbol{\beta}_0) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi^2(\mathbf{x}_i,y_i)}[\varphi^2(a_{in})-\varphi^2(F(z_i(\boldsymbol{\beta}_0)))]\nabla_{\boldsymbol{\beta}}g(\mathbf{x}_i,\boldsymbol{\beta}_0)\nabla_{\boldsymbol{\beta}}^{\tau}g(\mathbf{x}_i,\boldsymbol{\beta}_0)\\
&+ \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi^2(\mathbf{x}_i,y_i)}\varphi^2(F(z_i(\boldsymbol{\beta}_0)))\nabla_{\boldsymbol{\beta}}g(\mathbf{x}_i,\boldsymbol{\beta}_0)\nabla_{\boldsymbol{\beta}}^{\tau}g(\mathbf{x}_i,\boldsymbol{\beta}_0)\\
&= J_{1n}+J_{2n},
\end{aligned}$$

where

$$J_{1n} = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi^2(\mathbf{x}_i,y_i)}[\varphi^2(a_{in})-\varphi^2(F(z_i(\boldsymbol{\beta}_0)))]\nabla_{\boldsymbol{\beta}}g(\mathbf{x}_i,\boldsymbol{\beta}_0)\nabla_{\boldsymbol{\beta}}^{\tau}g(\mathbf{x}_i,\boldsymbol{\beta}_0)$$

and

$$J_{2n} = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi^2(\mathbf{x}_i,y_i)}\varphi^2(F(z_i(\boldsymbol{\beta}_0)))\nabla_{\boldsymbol{\beta}}g(\mathbf{x}_i,\boldsymbol{\beta}_0)\nabla_{\boldsymbol{\beta}}^{\tau}g(\mathbf{x}_i,\boldsymbol{\beta}_0).$$

Note that

$$\begin{aligned}
\|J_{1n}\| &\leq \frac{1}{n}\sum_{i=1}^{n}|\pi^{-2}(\mathbf{x}_i,y_i)|\|\nabla_{\boldsymbol{\beta}}g(\mathbf{x}_i,\boldsymbol{\beta}_0)\|^2|\varphi^2(a_{in})-\varphi^2(F(z_i(\boldsymbol{\beta}_0)))|\\
&\leq \left[\frac{1}{n}\sum_{i=1}^{n}|\pi^{-4}(\mathbf{x}_i,y_i)|\|\nabla_{\boldsymbol{\beta}}g(\mathbf{x}_i,\boldsymbol{\beta}_0)\|^4\right]^{1/2}\left[\max_{1\leq i\leq n}|\varphi^2(a_{in})-\varphi^2(F(z_i(\boldsymbol{\beta}_0)))|^2\right]^{1/2}.
\end{aligned}$$

By the continuity of φ and the fact that $a_{in} \xrightarrow{a.s.} F(z_i(\boldsymbol{\beta}_0))$ for all i (Hájek & Šidák, 1967), we have

$$\max_{1\leq i\leq n}|\varphi^2(a_{in})-\varphi^2(F(z_i(\boldsymbol{\beta}_0)))|^2 \xrightarrow{a.s.} 0.$$

Also, from the strong law of large numbers, we have

$$\frac{1}{n}\sum_{i=1}^{n}|\pi^{-4}(\mathbf{x}_i,y_i)|\|\nabla_{\boldsymbol{\beta}}g(\mathbf{x}_i,\boldsymbol{\beta}_0)\|^4 \xrightarrow{a.s.} E\big[\pi^{-4}(\mathbf{X},Y)\|\nabla_{\boldsymbol{\beta}}g(\mathbf{X},\boldsymbol{\beta}_0)\|^4\big] < \infty,$$

by assumption (I_2). Hence, $J_{1n} = o(1)$ with probability 1. When it comes to J_{2n}, a direct application of the strong law of large numbers yields

$$J_{2n} \xrightarrow{a.s.} E\big[\pi^{-1}(\mathbf{X},Y)\nabla_{\boldsymbol{\beta}}g(\mathbf{X},\boldsymbol{\beta}_0)\nabla_{\boldsymbol{\beta}}^{\tau}g(\mathbf{X},\boldsymbol{\beta}_0)\varphi^2(F(\varepsilon))\big] = A_{\boldsymbol{\beta}_0} \quad \text{by } (I_1),$$

as $E[\varphi^2(F(\varepsilon))|\mathbf{X}] = \int_0^1 \varphi^2(t)dt = 1$. Thus, $n^{-1}\sum_{i=1}^n v_i(\boldsymbol{\beta}_0)v_i^\tau(\boldsymbol{\beta}_0) \xrightarrow{a.s.} \mathbf{A}_{\boldsymbol{\beta}_0}$.

(iii) It is not difficult to see that $\sqrt{n}\widetilde{S}_n(\boldsymbol{\beta}_0)$ can be rewritten as

$$\begin{aligned}\sqrt{n}\widetilde{S}_n(\boldsymbol{\beta}_0) &= \sqrt{n}S_n(\boldsymbol{\beta}_0) + \frac{1}{\sqrt{n}}\sum_{i=1}^n\left\{\frac{\delta_i}{\widehat{\pi}(\mathbf{x}_i,y_i)} - \frac{\delta_i}{\pi(\mathbf{x}_i,y_i)}\right\}\nabla_{\boldsymbol{\beta}}g(\mathbf{x}_i,\boldsymbol{\beta}_0)\varphi\left(R(z_i(\boldsymbol{\beta}_0))/(n+1)\right)\\ &= \sqrt{n}S_n(\boldsymbol{\beta}_0) + J_{3n}.\end{aligned}$$

Following the same argument as in Niu et al. (2014), it can be obtained that

$$J_{3n} = \frac{1}{\sqrt{n}}\sum_{i=1}^n\left\{1 - \frac{\delta_i}{\pi(\mathbf{x}_i,y_i)}\right\}\nabla_{\boldsymbol{\beta}}g(\mathbf{x}_i,\boldsymbol{\beta}_0)E[\varphi(F(\varepsilon))|\,\mathbf{X},\delta=0] + o_p(1).$$

Similar arguments as in the proof of (i) give $E\{\sqrt{n}\widetilde{S}_n(\boldsymbol{\beta}_0)\} \to 0$ as $n\to\infty$. Putting $\mathbf{X}^\star = \nabla_{\boldsymbol{\beta}}g(\mathbf{X},\beta_0)$ and $\widetilde{\mathbf{B}}_{\boldsymbol{\beta}_0} = E\left[\pi^{-1}(\mathbf{X},Y)\mathbf{X}^\star\mathbf{X}^{\star\tau}\varphi^2(F(\varepsilon))\right] + E\{(\pi^{-1}(\mathbf{X},Y)-1)\mathbf{X}^\star\mathbf{X}^{\star\tau}E^2[\varphi(F(\varepsilon))|\mathbf{X},\delta=0]\}$, we have

$$\begin{aligned}\mathrm{Var}\{\sqrt{n}\widetilde{S}_n(\boldsymbol{\beta}_0)\} &= \widetilde{\mathbf{B}}_{\boldsymbol{\beta}_0} + 2E\left\{\frac{\delta}{\pi(\mathbf{X},Y)}\left[1-\frac{\delta}{\pi(\mathbf{X},Y)}\right]\mathbf{X}^\star\mathbf{X}^{\star\tau}\varphi(F(\varepsilon))E[\varphi(F(\varepsilon))|\mathbf{X},\delta=0]\right\}\\ &= \widetilde{\mathbf{B}}_{\boldsymbol{\beta}_0} + 2E\{(1-\pi^{-1}(\mathbf{X},Y))\mathbf{X}^\star\mathbf{X}^{\star\tau}E^2[\varphi(F(\varepsilon))|\mathbf{X},\delta=0]\} = \mathbf{B}_{\boldsymbol{\beta}_0}.\end{aligned}$$

From this, applying the argument in the proof of (ii), we have, $\sqrt{n}\widetilde{S}_n(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0},\mathbf{B}_{\boldsymbol{\beta}_0})$. On the other hand,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n\widetilde{v}_i(\boldsymbol{\beta}_0)\widetilde{v}_i(\boldsymbol{\beta}_0)^\tau &= \frac{1}{n}\sum_{i=1}^n\left(\frac{\delta_i}{\widehat{\pi}^2(\mathbf{x}_i,y_i)} - \frac{\delta_i}{\pi^2(\mathbf{x}_i,y_i)}\right)\nabla_{\boldsymbol{\beta}}g(\mathbf{x}_i,\boldsymbol{\beta}_0)\nabla_{\boldsymbol{\beta}}^\tau g(\mathbf{x}_i,\boldsymbol{\beta}_0)\varphi\left(\frac{R(z_i(\boldsymbol{\beta}_0)}{n+1}\right)\\ &+ \frac{1}{n}\sum_{i=1}^n\frac{\delta_i}{\pi^2(\mathbf{x}_i,y_i)}\nabla_{\boldsymbol{\beta}}g(\mathbf{x}_i,\boldsymbol{\beta}_0)\nabla_{\boldsymbol{\beta}}^\tau g(\mathbf{x}_i,\boldsymbol{\beta}_0)\varphi^2\left(\frac{R(z_i(\boldsymbol{\beta}_0)}{n+1}\right)\\ &= J_{4n} + J_{5n}.\end{aligned}$$

From the consistency of $\widehat{\pi}(\mathbf{x},y)$, $J_{4n} = o(1)\ a.s.$ and by the strong law of large number, $J_{5n} = \mathbf{A}_{\boldsymbol{\beta}_0} + o(1)\ a.s.$ Thus, $n^{-1}\sum_{i=1}^n\widetilde{v}_i(\boldsymbol{\beta}_0)\widetilde{v}_i(\boldsymbol{\beta}_0)^\tau \to \mathbf{A}_{\boldsymbol{\beta}_0}\ a.s.$ □

Proof of Theorem 1. Defining $T_n(\boldsymbol{\beta})$ as in equation (8.2), and following the same arguments as in the proof of Lemma 1 above, we have $\lim_{n\to\infty}\sup_{\boldsymbol{\beta}\in\mathscr{B}}\|S_n(\boldsymbol{\beta}) - T_n(\boldsymbol{\beta})\| = 0\ a.s.$ Thus, with probability 1, $S_n(\boldsymbol{\beta}) = T_n(\boldsymbol{\beta}) + o(1)$. From the fact that F is almost surely differentiable, so is $T_n(\boldsymbol{\beta})$. A Taylor expansion of $T_n(\boldsymbol{\beta})$ up to order 2 around $\boldsymbol{\beta}_0$ gives

$$T_n(\boldsymbol{\beta}) = T_n(\boldsymbol{\beta}_0) + [\nabla_{\boldsymbol{\beta}}T_n(\boldsymbol{\beta}_0)](\boldsymbol{\beta}-\boldsymbol{\beta}_0) + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\tau\nabla_{\boldsymbol{\beta}}^2T_n(\boldsymbol{\xi})(\boldsymbol{\beta}-\boldsymbol{\beta}_0), \tag{8.3}$$

where $\boldsymbol{\xi} = \lambda\boldsymbol{\beta}_0 + (1-\lambda)\boldsymbol{\beta}$, for some $\lambda\in(0,1)$. $\widehat{\boldsymbol{\beta}}_n$ being the solution of the estimation $S_n(\boldsymbol{\beta}) = 0$, yields

$$\mathbf{0} = T_n(\boldsymbol{\beta}_0) + [\nabla_{\boldsymbol{\beta}}T_n(\boldsymbol{\beta}_0)](\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0) + \frac{1}{2}(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)^\tau\nabla_{\boldsymbol{\beta}}^2T_n(\widehat{\boldsymbol{\xi}}_n)(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0) + o(1), \tag{8.4}$$

where $\widehat{\boldsymbol{\xi}}_n = \lambda\boldsymbol{\beta}_0 + (1-\lambda)\widehat{\boldsymbol{\beta}}_n$, for some $\lambda \in (0,1)$. Now from the boundedness of φ and derivatives of g by integrable functions independent of $\widetilde{\boldsymbol{\xi}}_n$, $\sqrt{n}\nabla^2_{\boldsymbol{\beta}} T_n(\widetilde{\boldsymbol{\xi}}_n)$ is almost surely bounded. Thus, from the strong consistency of $\widehat{\boldsymbol{\beta}}_n$, we have

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = [\nabla_{\boldsymbol{\beta}} T_n(\boldsymbol{\beta}_0)]^{-1}\sqrt{n}T_n(\boldsymbol{\beta}_0) + o(1). \tag{8.5}$$

On the other hand,

$$\begin{aligned}\nabla_{\boldsymbol{\beta}} T_n(\boldsymbol{\beta}_0) &= -\frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)}\nabla^\tau_{\boldsymbol{\beta}} g(\mathbf{x},\boldsymbol{\beta}_0)\nabla_{\boldsymbol{\beta}} g(\mathbf{x},\boldsymbol{\beta}_0) f(z_i(\boldsymbol{\beta}_0))\varphi'\big(F(z_i(\boldsymbol{\beta}_0))\big)\\ &\quad + \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)}\nabla^2_{\boldsymbol{\beta}} g(\mathbf{x},\boldsymbol{\beta}_0)\varphi\big(F(z_i(\boldsymbol{\beta}_0))\big),\end{aligned}$$

where $\nabla^2_{\boldsymbol{\beta}} g(\mathbf{x}_i, \boldsymbol{\beta}_0)$ is the Hessian matrix. Note that

$$\begin{aligned}E\left\{\frac{\delta_i}{\pi(\mathbf{X}_i, Y_i)}\nabla^2_{\boldsymbol{\beta}} g(\mathbf{X}_i,\boldsymbol{\beta}_0)\varphi\big(F(z_i(\boldsymbol{\beta}_0))\big)\right\} &= E\left\{E\left[\frac{\delta_i}{\pi(\mathbf{X}_i, Y_i)}\nabla^2_{\boldsymbol{\beta}} g(\mathbf{X}_i,\boldsymbol{\beta}_0)\varphi\big(F(z_i(\boldsymbol{\beta}_0))\big)\Big|\mathbf{X}_i\right]\right\}\\ &= E\left\{\frac{\delta_i}{\pi(\mathbf{X}_i, Y_i)}\nabla^2_{\boldsymbol{\beta}} g(\mathbf{X}_i,\boldsymbol{\beta}_0)E\left[\varphi\big(F(z_i(\boldsymbol{\beta}_0))\big)|\mathbf{X}_i\right]\right\}\end{aligned}$$

By assumption (I_1), $E\left[\varphi\big(F(z_i(\boldsymbol{\beta}_0))\big)|\mathbf{X}_i\right] = \int_0^1 \varphi(t)dt = 0$. The strong law of large numbers gives $n^{-1}\sum_{i=1}^n \big(\delta_i/\pi(\mathbf{x}_i, y_i)\big)\nabla^2_{\boldsymbol{\beta}} g(\mathbf{x},\boldsymbol{\beta}_0)\varphi\big(F(z_i(\boldsymbol{\beta}_0))\big) \xrightarrow{a.s.} E[\nabla^2_{\boldsymbol{\beta}} g(\mathbf{X},\boldsymbol{\beta}_0)\varphi(F(\varepsilon))] = 0$. Furthermore,

$$E[f(\varepsilon)\varphi'(F(\varepsilon))] = \int_{-\infty}^{\infty} f(\varepsilon)\varphi'(F(\varepsilon))dF(\varepsilon) = -\int_{-\infty}^{\infty} f'(\varepsilon)\varphi(F(\varepsilon))d\varepsilon,$$

from integration by parts, since $f(\varepsilon)\varphi(F(\varepsilon)) \to 0$ as $\varepsilon \to \pm\infty$. Now, putting $u = F(\varepsilon)$, we have

$$\int_{-\infty}^{\infty} f'(\varepsilon)\varphi(F(\varepsilon))d\varepsilon = -\int_0^1 \varphi(u)\varphi_f(u)du = -\gamma_\varphi^{-1},$$

as defined in Theorem 1. From this, applying the strong law of large numbers to $\nabla_{\boldsymbol{\beta}} T_n(\boldsymbol{\beta})$, gives $\nabla_{\boldsymbol{\beta}} T_n(\boldsymbol{\beta}) \to \gamma_\varphi^{-1}\mathbf{W}_{\boldsymbol{\beta}_0}$ $a.s.$, where $\mathbf{W}_{\boldsymbol{\beta}_0} = E[\nabla_{\boldsymbol{\beta}} g(\mathbf{X},\boldsymbol{\beta}_0)\nabla^\tau_{\boldsymbol{\beta}} g(\mathbf{X},\boldsymbol{\beta}_0)]$. This together with Equation (8.5), leads to

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, \gamma_\varphi^{-2}\mathbf{W}^{-1}_{\boldsymbol{\beta}_0}\mathbf{A}_{\boldsymbol{\beta}_0}\mathbf{W}^{-1}_{\boldsymbol{\beta}_0}).$$

Similarly, one can show that $\sqrt{n}(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, \gamma_\varphi^{-2}\mathbf{W}^{-1}_{\boldsymbol{\beta}_0}\mathbf{B}_{\boldsymbol{\beta}_0}\mathbf{W}^{-1}_{\boldsymbol{\beta}_0})$. Next, from the right hand side of (2.1), performing the Taylor expansion of $\log(\cdot)$ around 1, and substituting this one-term expansion into $L(\boldsymbol{\beta}_0, \gamma)$, there exists some ω_i between 1 and $1 + \boldsymbol{\xi}^\tau v_i(\boldsymbol{\beta}_0)$ such that

$$\begin{aligned}L(\boldsymbol{\beta}_0, \gamma) &= 2\sum_{i=1}^n \log\big(1 + \boldsymbol{\xi}^\tau v_i(\boldsymbol{\beta}_0)\big) = 2\sum_{i=1}^n \left[\boldsymbol{\xi}^\tau v_i(\boldsymbol{\beta}_0) - \frac{1}{3}\frac{\big(\boldsymbol{\xi}^\tau v_i(\boldsymbol{\beta}_0)\big)^3}{(1+\omega_i)^3}\right]\\ &= 2\boldsymbol{\xi}^\tau\sum_{i=1}^n v_i(\boldsymbol{\beta}_0) - \boldsymbol{\xi}^\tau\left\{\sum_{i=1}^n v_i(\boldsymbol{\beta}_0)v_i^\tau(\boldsymbol{\beta}_0)\right\}\boldsymbol{\xi} + O_p(n^{-1/2}).\end{aligned}$$

Taking the derivative with respect to $\boldsymbol{\xi}$ and setting it to $\mathbf{0}$, results in

$$\boldsymbol{\xi} = \left\{\frac{1}{n}\sum_{i=1}^{n} v_i(\boldsymbol{\beta}_0)v_i^{\tau}(\boldsymbol{\beta}_0)\right\}^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n} v_i(\boldsymbol{\beta}_0)\right\} + o_p(1).$$

$$\begin{aligned} L(\boldsymbol{\beta}_0,\gamma) &= \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} v_i(\boldsymbol{\beta}_0)\right\}^{\tau}\left\{\frac{1}{n}\sum_{i=1}^{n} v_i(\boldsymbol{\beta}_0)v_i^{\tau}(\boldsymbol{\beta}_0)\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} v_i(\boldsymbol{\beta}_0)\right\} + o_p(1) \\ &= \left\{\sqrt{n}S_n(\boldsymbol{\beta}_0)\right\}^{\tau}\left\{\frac{1}{n}\sum_{i=1}^{n} v_i(\boldsymbol{\beta}_0)v_i^{\tau}(\boldsymbol{\beta}_0)\right\}^{-1}\left\{\sqrt{n}S_n(\boldsymbol{\beta}_0)\right\} + o_p(1) \\ &= \left\{\sqrt{n}S_n(\boldsymbol{\beta}_0)\right\}^{\tau}\mathbf{A}_{\boldsymbol{\beta}_0}^{-1}\left\{\sqrt{n}S_n(\boldsymbol{\beta}_0)\right\} + o_p(1). \end{aligned}$$

By Lemma 1-(ii), we have $\sqrt{n}S_n(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(0,\mathbf{A}_{\boldsymbol{\beta}_0})$ as $n\to\infty$. $A_{\boldsymbol{\beta}_0}$ being positive definite by assumption (I_6), we have $\left\{\sqrt{n}S_n(\boldsymbol{\beta}_0)\right\}^{\tau}\mathbf{A}_{\boldsymbol{\beta}_0}^{-1}\left\{\sqrt{n}S_n(\boldsymbol{\beta}_0)\right\} \xrightarrow{\mathcal{D}} \chi_p^2$.

Similar arguments give

$$\begin{aligned} \tilde{L}(\boldsymbol{\beta}_0,\gamma) &= \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \widetilde{v}_i(\boldsymbol{\beta}_0)\right\}^{\tau}\left\{\frac{1}{n}\sum_{i=1}^{n} \widetilde{v}_i(\boldsymbol{\beta}_0)\widetilde{v}_i^{\tau}(\boldsymbol{\beta}_0)\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \widetilde{v}_i(\boldsymbol{\beta}_0)\right\} + o_p(1) \\ &= \left\{\sqrt{n}S_n(\boldsymbol{\beta}_0)\right\}^{\tau}\left\{\frac{1}{n}\sum_{i=1}^{n} \widetilde{v}_i(\boldsymbol{\beta}_0)\widetilde{v}_i^{\tau}(\boldsymbol{\beta}_0)\right\}^{-1}\left\{\sqrt{n}\widetilde{S}_n(\boldsymbol{\beta}_0)\right\} + o_p(1) \\ &= \left\{\sqrt{n}\widetilde{S}_n(\boldsymbol{\beta}_0)\right\}^{\tau}\mathbf{A}_{\boldsymbol{\beta}_0}^{-1}\left\{\sqrt{n}\widetilde{S}_n(\boldsymbol{\beta}_0)\right\} + o_p(1). \end{aligned}$$

By Lemma 1-(iii), we have $\sqrt{n}S_n(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(0,\mathbf{B}_{\boldsymbol{\beta}_0})$ as $n\to\infty$. Putting $\sqrt{n}S_n(\boldsymbol{\beta}_0) = \mathbf{B}_{\boldsymbol{\beta}_0}^{1/2}\mathbf{Z}$, where $\mathbf{Z}$ is the standard normal random p –vector, we have

$$\left(\sqrt{n}S_n(\boldsymbol{\beta}_0)\right)^{\tau}\mathbf{A}_{\boldsymbol{\beta}_0}^{-1}\left(\sqrt{n}S_n(\boldsymbol{\beta}_0)\right) = \mathbf{Z}^{\tau}\mathbf{B}_{\boldsymbol{\beta}_0}^{1/2}\mathbf{A}_{\boldsymbol{\beta}_0}^{-1}\mathbf{B}_{\boldsymbol{\beta}_0}^{1/2}\mathbf{Z} \xrightarrow{\mathcal{D}} \sum_{i=1}^{p}\lambda_i\chi_{1,i}^2,$$

where the λ_i are the eigenvalues of $\mathbf{B}_{\boldsymbol{\beta}_0}^{1/2}\mathbf{A}_{\boldsymbol{\beta}_0}^{-1}\mathbf{B}_{\boldsymbol{\beta}_0}^{1/2}$ and the $\chi_{1,i}^2$ are i.i.d. χ_1^2 random variables with one degree of freedom. Thus, the proof is complete. □

Proof of Theorem 3. Putting

$$B_{jn} = -\int(\hat{F}_{jn} - F_{jn})d\varphi(J_{jn}) + \int(\hat{J}_{jn} - J_{jn})\frac{dF_{jn}}{dJ_{jn}}d\varphi(J_{jn}),$$

it is shown in Brunner & Denker (1994) that $\boldsymbol{\Sigma}_{jn} = n^2\mathrm{Var}(B_{jn})$, as $U_{jn} = nB_{jn}$ in Lemma 2. Thus, our case corresponds to setting $M_0 = M_1 = 1$, $\alpha = 1$, and $m(n) = n$. Also, by definition,

$$S_n^j(\boldsymbol{\beta}_0) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\lambda}_i\varphi\left(\frac{R(\nu_{ij}(\boldsymbol{\beta}_0))}{n+1}\right) = \int\varphi\left(\frac{n}{n+1}\hat{J}_{jn}\right)dF_{jn}.$$

Now, from the fact that $\sigma^2(\varepsilon|\mathbf{x}) > 0$ and from assumption (I_6), there exists a positive constant c such that $\varsigma_{jn} \geq cn^2$ which satisfies the assumptions of Lemma 2, as φ is twice continuously differentiable with bounded

derivatives, and $\alpha < (a+1)/2$ with $a = 2$. By assumption (I_5), we have $E\{S_n^j(\boldsymbol{\beta}_0)\} \to \mathbf{0}$ as $n \to \infty$. Then, $n\boldsymbol{\Sigma}_{jn}^{-1}\Gamma_n^j(\boldsymbol{\beta}_0) = n\boldsymbol{\Sigma}_{jn}^{-1}S_n^j(\boldsymbol{\beta}_0) + o_p(1)$. By Lemma 2, $n\boldsymbol{\Sigma}_{jn}^{-1}\Gamma_n^j(\boldsymbol{\beta}_0)$ is asymptotically multivariate standard normal. Thus, we obtain

$$\sqrt{n}S_n^j(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, \Sigma_{\boldsymbol{\beta}_0}^j), \quad \text{where } \Sigma_{\boldsymbol{\beta}_0}^j = \lim_{n\to\infty} n^{-1}\boldsymbol{\Sigma}_{jn}\boldsymbol{\Sigma}_{jn}^{\tau}, \quad j = 1, 2.$$

Recall that $T_n^j(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{\lambda}_i\varphi\big(H_i^j(\zeta_{ij}(\boldsymbol{\beta}))\big)$. As in the proof of Theorem 1, taking into account the consistency of $\widehat{\pi}(\mathbf{x}, y)$ and $\widehat{m}_0(\mathbf{x}, \gamma)$, once again, we have $\lim_{n\to\infty}\sup_{\boldsymbol{\beta}\in\mathscr{B}}\|S_n^j(\boldsymbol{\beta}) - T_n^j(\boldsymbol{\beta})\| = 0\ a.s.$ Hence, with probability 1, $S_n^j(\boldsymbol{\beta}) = T_n^j(\boldsymbol{\beta}) + o(1)$. Now, performing a Taylor expansion of $T_n^j(\boldsymbol{\beta})$ up to order 2 around $\boldsymbol{\beta}_0$, we obtain

$$T_n^j(\boldsymbol{\beta}) = T_n^j(\boldsymbol{\beta}_0) + [\nabla_{\boldsymbol{\beta}}T_n^j(\boldsymbol{\beta}_0)](\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^{\tau}\nabla_{\boldsymbol{\beta}}^2 T_n^j(\boldsymbol{\xi})(\boldsymbol{\beta} - \boldsymbol{\beta}_0), \tag{8.6}$$

where $\boldsymbol{\xi} = \lambda\boldsymbol{\beta}_0 + (1-\lambda)\boldsymbol{\beta}$, for $\lambda \in (0, 1)$. Noting that $\widehat{\boldsymbol{\beta}}_n^j$ is a zero $S_n^j(\boldsymbol{\beta})$ and plugging $\widehat{\boldsymbol{\beta}}_n^j$ in equation (8.6), we get

$$\mathbf{0} = S_n^j(\boldsymbol{\beta}_0) + [\nabla_{\boldsymbol{\beta}}T_n^j(\boldsymbol{\beta}_0)](\widehat{\boldsymbol{\beta}}_n^j - \boldsymbol{\beta}_0) + \frac{1}{2}(\widehat{\boldsymbol{\beta}}_n^j - \boldsymbol{\beta}_0)^{\tau}\nabla_{\boldsymbol{\beta}}^2 T_n^j(\widehat{\boldsymbol{\xi}}_n^j)(\widehat{\boldsymbol{\beta}}_n^j - \boldsymbol{\beta}_0) + o(1), \tag{8.7}$$

where $\widehat{\boldsymbol{\xi}}_n^j = \lambda\boldsymbol{\beta}_0 + (1-\lambda)\widehat{\boldsymbol{\beta}}_n^j$, for $\lambda \in (0, 1)$. Now from the strong consistency of $\widehat{\boldsymbol{\beta}}_n^j$, and using assumptions $(I_1) - (I_6)$, the third term on the right hand side of equation (8.7) converges to 0 in probability. Therefore,

$$\sqrt{} \qquad\qquad \overline{n}S_n^j(\boldsymbol{\beta}_0) + o_p(1).$$

To this end,

$$\nabla_{\boldsymbol{\beta}}T_n^j(\boldsymbol{\beta}_0) \;=\; \frac{1}{n}\sum_{i=1}^n \boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^{\tau}h_i^j(\zeta_{ij}(\boldsymbol{\beta}_0))\varphi'\big(H_i^j(\zeta_{ij}(\boldsymbol{\beta}_0))\big) + \frac{1}{n}\sum_{i=1}^n \nabla_{\boldsymbol{\beta}}^2 g(\mathbf{x}_i, \boldsymbol{\beta}_0)\varphi\big(H_i^j(\zeta_{ij}(\boldsymbol{\beta}_0))\big).$$

From strong law of large numbers, $n^{-1}\sum_{i=1}^n \boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^{\tau}h_i^j(\zeta_{ij}(\boldsymbol{\beta}_0))\varphi'\big(H_i^j(\zeta_{ij}(\boldsymbol{\beta}_0))\big)$ converges almost surely to $E\left\{\nabla_{\boldsymbol{\beta}}g(\mathbf{X}, \boldsymbol{\beta}_0)\nabla_{\boldsymbol{\beta}}^{\tau}g(\mathbf{X}, \boldsymbol{\beta}_0)h^j(\zeta_j(\boldsymbol{\beta}_0))\varphi'\big(H^j(\zeta_j(\boldsymbol{\beta}_0))\big)\right\}$. Also, we have that

$$\frac{1}{n}\sum_{i=1}^n \nabla_{\boldsymbol{\beta}}^2 g(\mathbf{x}_i, \boldsymbol{\beta}_0)\varphi\big(H_i^j(\zeta_{ij}(\boldsymbol{\beta}_0))\big) \to E\{\nabla_{\boldsymbol{\beta}}^2 g(\mathbf{X}, \boldsymbol{\beta}_0)\varphi\big(H^j(\zeta_j(\boldsymbol{\beta}_0))\big)\}\ a.s.$$

Thus, with probability 1, $\lim_{n\to\infty}\nabla_{\boldsymbol{\beta}}T_n^j(\boldsymbol{\beta}_0) = \mathbf{V}_j$. From the fact that $\sqrt{n}S_n^j(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, \Sigma_{\boldsymbol{\beta}_0}^j)$, we have

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, \mathbf{M}_j), \quad \text{where } \mathbf{M}_j = \mathbf{V}_j^{-1}\Sigma_{\boldsymbol{\beta}_0}^j\mathbf{V}_j^{-1}.$$

□

Proof of Theorem 5. In this proof, M is taken to be a positive constant not necessarily the same. Recall from equation (3.9) that the log likelihood ratio of $\boldsymbol{\beta}_0$ is given by

$$-2\log R_n^j(\boldsymbol{\beta}_0) = -2\log\prod_{i=1}^n\big(1 + \boldsymbol{\xi}^{\tau}\eta_{ij}(\boldsymbol{\beta}_0)\big)^{-1} = 2\sum_{i=1}^n\log\big(1 + \boldsymbol{\xi}^{\tau}\eta_{ij}(\boldsymbol{\beta}_0)\big).$$

By $(I_1)-(I_4)$, there exist a positive constant M and a function $h \in L^p$, $p \geq 1$ such that $|\varphi(t)| \leq M$ for all $t \in (0,1)$, and $\|\nabla_{\boldsymbol{\beta}} g(\mathbf{x}_i, \boldsymbol{\beta}_0)\| \leq h(\mathbf{x}_i)$, where $\|\cdot\|$ stands for the L^2- norm. Since $E(|h(\mathbf{x}_i)|^p) < \infty$ for $p \geq 1$, we have $\max_{1\leq i\leq n} \|\nabla_{\boldsymbol{\beta}} g(\mathbf{x}_i, \boldsymbol{\beta}_0)\| = o_p(n^{1/2})$. Also, $\|\eta_{ij}(\boldsymbol{\beta}_0)\| \leq M \times \max_{1\leq i\leq n} h(\mathbf{x}_i)$, which implies that

$$\max_{1\leq i\leq n} \|\eta_{ij}(\boldsymbol{\beta}_0)\| = o_p(n^{1/2}) \quad and \quad \frac{1}{n}\sum_{i=1}^{n} \|\eta_{ij}(\boldsymbol{\beta}_0)\|^3 = o_p(n^{1/2}). \tag{8.8}$$

Putting $\boldsymbol{\Lambda}_{nj} = \text{Var}(\sqrt{n} S_n^j(\boldsymbol{\beta}_0))$, we have $\boldsymbol{\Lambda}_{nj} = \Sigma^j_{\boldsymbol{\beta}_0} + o_p(1)$, from which $\Sigma^j_{\boldsymbol{\beta}_0}$ is assumed to be positive definite. Also, we can show that $n^{-1}\sum_{i=1}^n \eta_{ij}(\boldsymbol{\beta}_0)\eta_{ij}^\tau(\boldsymbol{\beta}_0) - \boldsymbol{\Lambda}_{nj} = o_p(1)$ for $j = 1, 2$. Since $\sqrt{n} S_n^j(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N(0, \Sigma^j_{\boldsymbol{\beta}_0})$, we have $\|S_n^j(\boldsymbol{\beta}_0)\| = O_p(n^{-1/2})$. Now from equation (8.8), using the same argument as in Owen (1990), $\|\boldsymbol{\xi}\| = O_p(n^{-1/2})$. Following similar arguments as in Theorem 1,

$$\begin{aligned} -2\log R_n^j(\boldsymbol{\beta}_0) &= \sum_{i=1}^n \boldsymbol{\xi}^\tau \eta_{ij}(\boldsymbol{\beta}_0) + o_p(1) \\ &= \left\{ \frac{1}{} \qquad\qquad\qquad\qquad \frac{}{n}\sum_{i=1}^n \eta_{ij}(\boldsymbol{\beta}_0) \right\} + o_p(1) \\ &= \left\{ \sqrt{n}\boldsymbol{\Lambda}_{nj}^{-1/2} S_n^j(\boldsymbol{\beta}_0) \right\}^\tau \left\{ \sqrt{n}\boldsymbol{\Lambda}_{nj}^{-1/2} S_n^j(\boldsymbol{\beta}_0) \right\} + o_p(1). \end{aligned}$$

Using Slutsky's lemma, we have $\sqrt{n}\boldsymbol{\Lambda}_{nj}^{-1/2} S_n^j(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(0, I_p)$ as $n \to \infty$, and therefore,

$$-2\log R_n^j(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} \chi_p^2.$$

□

Table 1: 95% coverage probabilities (Average lengths of 95% confidence intervals) of β_2 for the linear model under t_{df} with $n = 200$ and regression simple imputation (SI).

cases	df	SI-NA$_{LS}$	SI-EL$_{LS}$	SI-NA$_R$	SI-EL$_R$	WNA$_R$	WEL$_R$
	5	96.76% (0.92)	96.06% (0.68)	96.16% (0.78)	95.09% (0.43)	95.98% (0.73)	94.99% (0.39)
	15	95.95% (0.90)	95.67% (0.62)	95.69% (0.71)	95.05% (0.37)	95.59% (0.65)	94.97% (0.33)
case 1	25	95.78% (0.87)	95.48% (0.57)	95.51% (0.68)	95.00% (0.32)	95.51% (0.61)	95.01% (0.27)
	40	95.68% (0.77)	95.35% (0.53)	95.18% (0.64)	95.01% (0.28)	95.34% (0.55)	94.98% (0.25)
	50	95.13% (0.63)	95.08% (0.44)	95.10% (0.56)	94.99% (0.24)	95.05% (0.51)	94.94% (0.20)
	5	96.86% (0.97)	96.10% (0.70)	96.21% (0.77)	95.06% (0.45)	95.96% (0.72)	95.08% (0.40)
	15	95.93% (0.93)	95.91% (0.63)	95.78% (0.72)	95.04% (0.41)	95.73% (0.66)	95.00% (0.35)
case 2	25	95.57% (0.85)	95.43% (0.59)	95.53% (0.65)	95.00% (0.36)	95.49% (0.58)	94.97% (0.30)
	40	95.38% (0.78)	95.22% (0.54)	95.31% (0.61)	94.99% (0.32)	95.29% (0.53)	94.95% (0.26)
	50	95.27% (0.69)	95.14% (0.46)	95.18% (0.57)	95.01% (0.27)	95.16% (0.50)	94.94% (0.21)
	5	96.96% (1.22)	96.16% (0.90)	96.27% (0.97)	95.13% (0.79)	96.08% (0.93)	95.17% (0.74)
	15	96.31% (0.97)	95.97% (0.77)	95.89% (0.85)	95.06% (0.73)	95.97% (0.81)	95.01% (0.63)
case 3	25	95.86% (0.89)	95.72% (0.68)	95.68% (0.73)	95.01% (0.68)	95.81% (0.69)	94.99% (0.60)
	40	95.57% (0.81)	95.43% (0.61)	95.51% (0.69)	94.97% (0.54)	95.42% (0.64)	94.97% (0.48)
	50	95.41% (0.76)	95.25% (0.57)	95.29% (0.64)	95.00% (0.49)	95.23% (0.60)	95.02% (0.40)
	5	96.49% (0.94)	96.03% (0.73)	96.07% (0.78)	95.07% (0.44)	95.94% (0.75)	95.31% (0.43)
	15	95.87% (0.87)	95.66% (0.65)	95.72% (0.70)	95.06% (0.39)	95.63% (0.64)	95.07% (0.38)
case 4	25	95.69% (0.75)	95.28% (0.59)	95.34% (0.67)	95.01% (0.34)	95.28% (0.60)	95.01% (0.32)
	40	95.23% (0.64)	95.17% (0.56)	95.21% (0.64)	94.97% (0.30)	95.14% (0.52)	94.98% (0.28)
	50	95.13% (0.62)	95.05% (0.47)	95.08% (0.57)	94.99% (0.25)	95.03% (0.49)	94.93% (0.22)
	5	96.47% (0.91)	96.08% (0.71)	96.09% (0.75)	94.99% (0.45)	95.97% (0.71)	95.04% (0.39)
	15	95.76% (0.83)	95.57% (0.67)	95.97% (0.72)	95.02% (0.41)	96.82% (0.64)	94.99% (0.34)
case 5	25	95.46% (0.79)	95.28% (0.60)	95.46% (0.67)	94.88% (0.39)	95.30% (0.60)	95.00% (0.29)
	40	95.33% (0.61)	95.12% (0.55)	95.13% (0.61)	95.00% (0.34)	95.10% (0.54)	95.01% (0.25)
	50	95.07% (0.56)	95.03% (0.48)	95.04% (0.54)	94.98% (0.28)	95.03% (0.50)	94.98% (0.21)

Table 2: 95% coverage probabilities (Average lengths of 95% confidence intervals) of β_2 for the linear model under t_{df} with $n = 200$ and the weighted inverse marginal probability regression imputation (IP).

cases	df	IP-NA$_{LS}$	IP-EL$_{LS}$	IP-NA$_R$	IP-EL$_R$	WNA$_R$	WEL$_R$
	5	96.78% (0.85)	96.01% (0.61)	96.04% (0.70)	95.03% (0.33)	95.76% (0.69)	95.00% (0.27)
	15	95.75% (0.78)	95.37% (0.56)	95.51% (0.62)	95.01% (0.24)	95.43% (0.60)	94.99% (0.24)
case 1	25	95.48% (0.71)	95.23% (0.51)	95.25% (0.56)	94.98% (0.22)	95.31% (0.55)	94.96% (0.19)
	40	95.27% (0.58)	95.14% (0.45)	95.15% (0.51)	94.99% (0.20)	95.22% (0.49)	95.01% (0.17)
	50	95.12% (0.49)	95.07% (0.39)	95.10% (0.47)	95.02% (0.18)	95.11% (0.43)	94.98% (0.14)
	5	96.76% (0.87)	96.03% (0.69)	96.08% (0.72)	95.13% (0.31)	95.85% (0.67)	95.04% (0.30)
	15	95.84% (0.76)	95.35% (0.52)	95.63% (0.64)	95.06% (0.27)	95.59% (0.56)	95.02% (0.25)
case 2	25	95.57% (0.70)	95.21% (0.43)	95.29% (0.53)	94.94% (0.24)	95.37% (0.44)	94.99% (0.21)
	40	95.23% (0.53)	95.11% (0.40)	95.20% (0.48)	94.97% (0.21)	95.24% (0.41)	94.99% (0.19)
	50	95.08% (0.46)	95.05% (0.35)	95.09% (0.43)	95.03% (0.19)	95.12% (0.38)	95.01% (0.16)
	5	96.86% (0.98)	96.09% (0.87)	96.24% (0.94)	95.37% (0.70)	95.98% (0.83)	95.09% (0.61)
	15	96.21% (0.87)	95.77% (0.79)	95.83% (0.83)	95.02% (0.63)	95.70% (0.74)	95.00% (0.55)
case 3	25	95.67% (0.80)	95.27% (0.71)	95.35% (0.76)	94.96% (0.59)	95.41% (0.65)	94.98% (0.48)
	40	95.29% (0.73)	95.18% (0.60)	95.23% (0.68)	94.89% (0.51)	95.32% (0.61)	94.99% (0.42)
	50	95.18% (0.65)	95.12% (0.55)	95.15% (0.62)	95.01% (0.43)	95.09% (0.57)	95.03% (0.35)
	5	96.52% (0.82)	96.07% (0.61)	96.03% (0.69)	95.08% (0.32)	95.83% (0.63)	95.03% (0.28)
	15	95.79% (0.75)	95.69% (0.52)	95.62% (0.60)	94.98% (0.29)	95.61% (0.54)	95.00% (0.26)
case 4	25	95.46% (0.68)	95.33% (0.42)	95.41% (0.53)	95.02% (0.23)	95.39% (0.46)	94.99% (0.22)
	40	95.19% (0.51)	95.13% (0.40)	95.20% (0.49)	94.94% (0.19)	95.20% (0.43)	95.01% (0.19)
	50	95.04% (0.45)	94.99% (0.36)	95.07% (0.43)	94.99% (0.17)	95.07% (0.39)	94.98% (0.17)
	5	96.68% (0.84)	96.04% (0.65)	96.14% (0.71)	95.05% (0.33)	95.87% (0.64)	94.99% (0.29)
	15	95.66% (0.73)	95.33% (0.59)	95.71% (0.69)	95.01% (0.28)	95.52% (0.59)	95.02% (0.25)
case 5	25	95.37% (0.67)	95.24% (0.50)	95.39% (0.57)	94.97% (0.24)	95.25% (0.51)	95.04% (0.23)
	40	95.20% (0.50)	95.14% (0.44)	95.25% (0.51)	94.98% (0.20)	95.13% (0.45)	95.01% (0.18)
	50	95.06% (0.43)	95.03% (0.38)	95.10% (0.47)	95.03% (0.18)	95.04% (0.41)	94.95% (0.16)

Table 3: 95% coverage probabilities (Average lengths of 95% confidence intervals) of β_2 for the linear model under $\mathcal{CN}(\epsilon)$ with $n = 200$ and regression simple imputation (SI).

cases	ϵ	SI-NA$_{LS}$	SI-EL$_{LS}$	SI-NA$_R$	SI-EL$_R$	WNA$_R$	WEL$_R$
	0.00	95.06% (0.164)	94.98% (0.133)	95.12% (0.145)	95.07% (0.134)	95.03% (0.142)	95.02% (0.121)
case 1	0.30	95.52% (0.253)	95.28% (0.209)	95.36% (0.215)	95.03% (0.168)	95.54% (0.204)	94.89% (0.152)
	0.50	95.76% (0.318)	95.46% (0.291)	95.47% (0.301)	95.05% (0.218)	95.75% (0.292)	94.94% (0.205)
	0.00	95.03% (0.163)	95.01% (0.131)	95.09% (0.146)	95.07% (0.137)	95.09% (0.143)	95.03% (0.123)
case 2	0.30	95.61% (0.249)	95.24% (0.203)	95.33% (0.211)	94.99% (0.159)	95.37% (0.207)	95.01% (0.155)
	0.50	95.87% (0.313)	95.69% (0.279)	95.73% (0.307)	95.03% (0.227)	95.78% (0.301)	94.99% (0.209)
	0.00	95.09% (0.203)	95.04% (0.197)	95.23% (0.199)	95.03% (0.189)	95.11% (0.203)	95.02% (0.181)
case 3	0.30	95.73% (0.298)	95.55% (0.231)	95.44% (0.238)	94.93% (0.207)	95.42% (0.232)	95.05% (0.201)
	0.50	95.99% (0.389)	95.61% (0.347)	95.49% (0.353)	95.02% (0.277)	95.50% (0.343)	94.97% (0.271)
	0.00	95.07% (0.161)	95.02% (0.130)	95.11% (0.142)	95.05% (0.135)	95.08% (0.140)	95.01% (0.119)
case 4	0.30	95.56% (0.236)	95.21% (0.200)	95.32% (0.208)	95.00% (0.153)	95.28% (0.201)	94.98% (0.143)
	0.50	95.73% (0.319)	95.56% (0.278)	95.38% (0.285)	94.98% (0.224)	95.52% (0.277)	95.00% (0.207)
	0.00	95.05% (0.167)	94.99% (0.134)	95.11% (0.151)	95.06% (0.131)	95.09% (0.146)	95.04% (0.127)
case 5	0.30	95.67% (0.254)	95.29% (0.213)	95.44% (0.215)	95.04% (0.169)	95.43% (0.205)	95.01% (0.163)
	0.50	95.78% (0.328)	95.60% (0.301)	95.68% (0.316)	94.99% (0.229)	95.62% (0.304)	94.98% (0.213)

Table 4: 95% coverage probabilities (Average lengths of 95% confidence intervals) of β_2 for the linear model under $\mathcal{CN}(\epsilon)$ with $n = 200$ and the weighted inverse marginal probability regression imputation (IP).

cases	ϵ	IP-NA$_{LS}$	IP-EL$_{LS}$	IP-NA$_R$	IP-EL$_R$	WNA$_R$	WEL$_R$
	0.00	95.03% (0.113)	94.98% (0.098)	95.08% (0.111)	95.04% (0.071)	95.07% (0.103)	95.01% (0.065)
case 1	0.30	95.25% (0.163)	95.16% (0.127)	96.19% (0.157)	94.97% (0.117)	95.19% (0.150)	94.99% (0.109)
	0.50	95.49% (0.198)	95.23% (0.143)	95.32% (0.166)	95.05% (0.125)	95.41% (0.159)	95.02% (0.119)
	0.00	94.94% (0.112)	95.01% (0.096)	95.12% (0.110)	95.01% (0.069)	95.10% (0.105)	94.99% (0.067)
case 2	0.30	95.37% (0.171)	95.21% (0.136)	95.29% (0.162)	95.00% (0.124)	95.21% (0.156)	94.95% (0.117)
	0.50	95.68% (0.201)	95.49% (0.177)	95.59% (0.198)	95.03% (0.133)	95.52% (0.184)	94.98% (0.123)
	0.00	95.14% (0.171)	95.06% (0.126)	95.13% (0.135)	95.05% (0.113)	95.12% (0.134)	95.01% (0.108)
case 3	0.30	95.63% (0.204)	95.34% (0.179)	95.46% (0.187)	95.03% (0.128)	95.53% (0.181)	94.98% (0.122)
	0.50	95.87% (0.245)	95.56% (0.195)	95.67% (0.212)	94.99% (0.143)	95.61% (0.209)	95.03% (0.139)
	0.00	95.06% (0.111)	94.99% (0.093)	95.11% (0.102)	95.00% (0.067)	95.08% (0.101)	95.06% (0.062)
case 4	0.30	95.23% (0.155)	95.10% (0.123)	95.18% (0.131)	95.08% (0.109)	95.23% (0.129)	95.01% (0.105)
	0.50	95.47% (0.183)	95.27% (0.134)	95.39% (0.157)	94.97% (0.121)	95.36% (0.148)	95.02% (0.117)
	0.00	95.03% (0.116)	95.01% (0.097)	96.07% (0.113)	95.04% (0.073)	95.07% (0.106)	95.00% (0.067)
case 5	0.30	95.28% (0.175)	95.19% (0.143)	95.23% (0.169)	95.07% (0.129)	95.21% (0.159)	94.93% (0.121)
	0.50	95.65% (0.200)	95.42% (0.151)	95.53% (0.186)	95.01% (0.135)	95.48% (0.173)	95.02% (0.132)

Table 5: 95% coverage probabilities (Average lengths of 95% confidence intervals) of β_2 for the linear model under the Laplace distribution and regression simple imputation (SI).

cases	n	SI-NA$_{LS}$	SI-EL$_{LS}$	SI-NA$_R$	SI-EL$_R$	WNA$_R$	WEL$_R$
	15	96.17% (1.305)	95.98% (1.008)	96.09% (1.115)	95.16% (0.897)	95.97% (1.006)	95.07% (0.881)
case 1	50	95.83% (0.997)	95.45% (0.899)	95.57% (0.912)	95.08% (0.789)	95.83% (0.901)	95.03% (0.754)
	100	95.59% (0.791)	95.28% (0.513)	95.46% (0.553)	95.01% (0.376)	95.43% (0.542)	94.98% (0.354)
	250	95.01% (0.221)	94.98% (0.187)	95.03% (0.195)	94.99% (0.133)	95.01% (0.189)	94.97% (0.125)
	15	96.28% (1.312)	96.01% (1.083)	96.13% (1.128)	95.13% (0.889)	95.95% (1.004)	95.02% (0.877)
case 2	50	95.91% (1.064)	95.67% (0.937)	95.78% (0.997)	94.99% (0.827)	95.71% (0.978)	94.97% (0.819)
	100	95.73% (0.887)	95.39% (0.752)	95.51% (0.831)	95.05% (0.387)	95.47% (0.816)	95.01% (0.379)
	250	94.97% (0.237)	95.03% (0.198)	94.96% (0.205)	95.00% (0.148)	95.02% (0.199)	95.01% (0.135)
	15	96.49% (1.433)	96.16% (1.119)	96.27% (1.217)	95.11% (1.091)	96.02% (1.108)	95.07% (1.009)
case 3	50	96.07% (1.202)	95.87% (1.007)	95.94% (1.098)	95.08% (0.913)	95.91% (1.003)	95.03% (0.904)
	100	95.58% (0.989)	95.37% (0.863)	95.42% (0.873)	94.98% (0.718)	95.32% (0.856)	94.99% (0.708)
	250	95.13% (0.803)	94.89% (0.721)	95.08% (0.735)	95.01% (0.505)	95.05% (0.723)	94.95% (0.499)
	15	96.13% (1.267)	95.75% (0.995)	96.04% (1.102)	95.14% (0.846)	95.97% (1.095)	95.02% (0.839)
case 4	50	95.79% (0.953)	95.38% (0.872)	95.53% (0.901)	94.99% (0.752)	95.46% (0.899)	95.00% (0.738)
	100	95.47% (0.725)	95.21% (0.501)	95.29% (0.698)	95.03% (0.302)	95.22% (0.637)	94.98% (0.298)
	250	94.88% (0.202)	94.96% (0.173)	95.00% (0.187)	94.99% (0.125)	95.03% (0.181)	94.97% (0.121)
	15	96.15% (1.299)	95.88% (1.003)	96.05% (1.103)	95.10% (0.837)	95.98% (1.003)	95.04% (0.825)
case 5	50	95.74% (0.913)	95.44% (0.855)	95.62% (0.879)	95.04% (0.748)	95.51% (0.867)	95.02% (0.733)
	100	95.21% (0.778)	95.13% (0.503)	95.18% (0.688)	95.02% (0.295)	95.13% (0.678)	94.99% (0.275)
	250	95.03% (0.199)	94.97% (0.175)	94.99% (0.191)	95.01% (0.123)	95.03% (0.183)	94.97% (0.117)

Table 6: 95% coverage probabilities (Average lengths of 95% confidence intervals) of β_2 for the linear model under the Laplace distribution and the weighted inverse marginal probability regression imputation (IP).

cases	n	IP-NA$_{LS}$	IP-EL$_{LS}$	IP-NA$_R$	IP-EL$_R$	WNA$_R$	WEL$_R$
	15	96.05% (1.041)	95.65% (0.916)	96.03% (0.989)	95.05% (0.725)	95.92% (0.975)	95.01% (0.698)
case 1	50	95.55% (0.872)	95.24% (0.742)	95.29% (0.796)	95.02% (0.521)	95.23% (0.783)	94.99% (0.487)
	100	95.24% (0.725)	95.13% (0.621)	95.17% (0.684)	95.01% (0.225)	95.12% (0.667)	95.02% (0.214)
	250	95.04% (0.184)	95.00% (0.131)	94.91% (0.148)	95.03% (0.096)	95.01% (0.139)	94.97% (0.088)
	15	96.08% (1.078)	95.71% (0.923)	96.04% (0.999)	95.09% (0.755)	95.95% (0.969)	95.03% (0.702)
case 2	50	95.63% (0.894)	95.31% (0.802)	95.41% (0.832)	95.07% (0.601)	95.31% (0.787)	94.98% (0.493)
	100	95.33% (0.746)	95.09% (0.695)	95.19% (0.709)	95.02% (0.279)	95.17% (0.671)	95.01% (0.225)
	250	94.88% (0.215)	94.90% (0.162)	94.98% (0.173)	95.02% (0.111)	94.99% (0.145)	95.05% (0.093)
	15	96.38% (1.223)	95.88% (1.009)	96.12% (1.111)	95.01% (0.967)	96.01% (1.081)	95.04% (0.959)
case 3	50	95.83% (0.989)	95.47% (0.836)	95.61% (0.857)	95.02% (0.773)	95.49% (0.833)	95.00% (0.764)
	100	95.51% (0.876)	95.26% (0.735)	95.31% (0.768)	94.99% (0.612)	95.25% (0.737)	94.96% (0.609)
	250	95.15% (0.559)	95.06% (0.478)	94.94% (0.483)	94.98% (0.309)	95.03% (0.473)	94.97% (0.301)
	15	96.09% (1.005)	95.84% (0.926)	96.05% (0.939)	94.98% (0.727)	95.93% (0.967)	95.03% (0.722)
case 4	50	95.65% (0.901)	95.23% (0.731)	95.38% (0.747)	95.00% (0.502)	95.30% (0.742)	95.01% (0.497)
	100	95.19% (0.692)	95.07% (0.491)	95.16% (0.503)	95.01% (0.267)	95.09% (0.489)	94.99% (0.263)
	250	94.92% (0.187)	94.99% (0.128)	94.93% (0.131)	94.95% (0.087)	94.98% (0.124)	95.01% (0.082)
	15	96.03% (0.999)	95.73% (0.912)	96.05% (0.925)	95.09% (0.743)	95.91% (0.969)	95.02% (0.734)
case 5	50	95.57% (0.875)	95.28% (0.738)	95.43% (0.761)	95.02% (0.501)	95.32% (0.745)	95.00% (0.499)
	100	95.18% (0.655)	95.10% (0.484)	95.14% (0.495)	95.00% (0.278)	95.10% (0.469)	95.01% (0.258)
	250	94.97% (0.183)	95.03% (0.119)	94.96% (0.125)	94.78% (0.083)	94.96% (0.119)	94.88% (0.0.79)

Table 7: 95% coverage probabilities (Average lengths of 95% confidence intervals) of β for the nonlinear Micheaelis-Menten model under t_3 and $\mathcal{CN}(0.9)$ with $n = 150$ and regression simple imputation (SI)

cases		SI-NA$_{LS}$	SI-EL$_{LS}$	SI-NA$_R$	SI-EL$_R$
case 1	$\mathcal{CN}(0.9)$	97.13% (3.76)	95.57% (2.19)	96.32% (2.78)	95.20% (1.43)
	t_3	97.35% (3.93)	95.43% (2.33)	96.67% (2.84)	95.15% (1.55)
case 2	$\mathcal{CN}(0.9)$	97.17% (3.73)	95.46% (2.16)	96.26% (2.67)	95.13% (1.37)
	t_3	97.26% (3.88)	95.47% (2.27)	96.51% (2.71)	95.11% (1.47)
case 3	$\mathcal{CN}(0.9)$	97.96% (4.12)	95.78% (3.05)	96.38% (3.47)	95.39% (2.23)
	t_3	97.98% (4.17)	95.69% (3.16)	96.29% (3.51)	95.27% (2.27)
case 4	$\mathcal{CN}(0.9)$	97.24% (3.72)	95.66% (2.28)	96.25% (2.81)	95.16% (1.51)
	t_3	97.38% (3.79)	95.72% (2.31)	96.33% (2.87)	95.21% (1.67)
case 5	$\mathcal{CN}(0.9)$	97.16% (3.52)	95.36% (2.11)	96.16% (2.61)	95.09% (134)
	t_3	97.23% (3.65)	95.41% (2.24)	95.69% (2.73)	95.05% (1.42)

Table 8: 95% coverage probabilities (Average lengths of 95% confidence intervals) of β for the nonlinear Micheaelis-Menten model under t_3 and $\mathcal{CN}(0.9)$ with $n = 150$ and the weighted inverse marginal probability regression imputation (IP)

cases		IP-NA$_{LS}$	IP-EL$_{LS}$	IP-NA$_R$	IP-EL$_R$	WNA$_R$	WEL$_R$
case 1	$\mathcal{CN}(0.9)$	96.96% (2.92)	95.16% (1.98)	95.97% (2.03)	95.09% (1.13)	95.43% (2.01)	94.99% (1.08)
	t_3	97.05% (2.98)	95.23% (2.02)	95.99% (2.07)	95.05% (1.17)	95.64% (2.04)	94.97% (1.10)
case 2	$\mathcal{CN}(0.9)$	96.93% (2.89)	95.09% (1.93)	95.86% (1.99)	95.03% (1.09)	95.37% (1.87)	95.02% (1.04)
	t_3	96.97% (2.95)	95.17% (1.99)	95.91% (2.01)	95.06% (1.13)	95.55% (1.92)	94.97% (1.07)
case 3	$\mathcal{CN}(0.9)$	97.16% (3.25)	95.36% (2.68)	96.23% (2.78)	95.13% (1.73)	95.87% (2.71)	95.07% (1.41)
	t_3	97.24% (3.33)	95.47% (2.73)	96.39% (2.81)	95.17% (1.87)	95.94% (2.79)	95.11% (1.53)
case 4	$\mathcal{CN}(0.9)$	96.89% (2.88)	95.10% (1.89)	95.76% (1.97)	95.05% (1.06)	95.29% (1.91)	95.05% (1.03)
	t_3	96.94% (2.93)	95.12% (1.92)	95.89% (1.99)	95.08% (1.11)	95.47% (1.96)	94.95% (1.05)
case 5	$\mathcal{CN}(0.9)$	96.86% (2.85)	95.06% (1.78)	95.68% (1.89)	95.02% (1.04)	95.53% (1.85)	94.98% (1.01)
	t_3	96.93% (2.91)	95.08% (1.87)	95.73% (1.96)	95.05% (1.09)	95.67% (1.92)	95.02% (1.03)

Table 9: lengths of 95% confidence intervals for the regression parameters for the Baby Weights data with missing rate 43%

		CC		WCC		IP		SI	
Method	Variable	NA	EL	NA	EL	NA	EL	NA	EL
	BMI	0.127	0.075	0.070	0.031	0.073	0.034	0.099	0.051
	Hispanic	0.093	0.036	0.039	0.015	0.046	0.023	0.063	0.038
LS	Smoke	3.583	1.716	1.980	1.270	2.080	1.321	2.838	1.769
	Race	1.970	0.973	1.050	0.789	1.150	0.873	1.569	0.974
	Age	1.942	0.867	1.007	0.701	1.131	0.771	1.544	0.783
	BMI	0.135	0.034	0.045	0.013	0.048	0.014	0.051	0.029
	Hispanic	0.099	0.021	0.019	0.009	0.025	0.013	0.037	0.021
R	Smoke	3.834	1.145	1.305	0.963	1.335	1.073	1.817	1.176
	Race	2.108	0.685	0.829	0.498	0.928	0.547	1.075	0.655
	Race	2.078	0.546	0.817	0.393	0.926	0.436	1.062	0.447

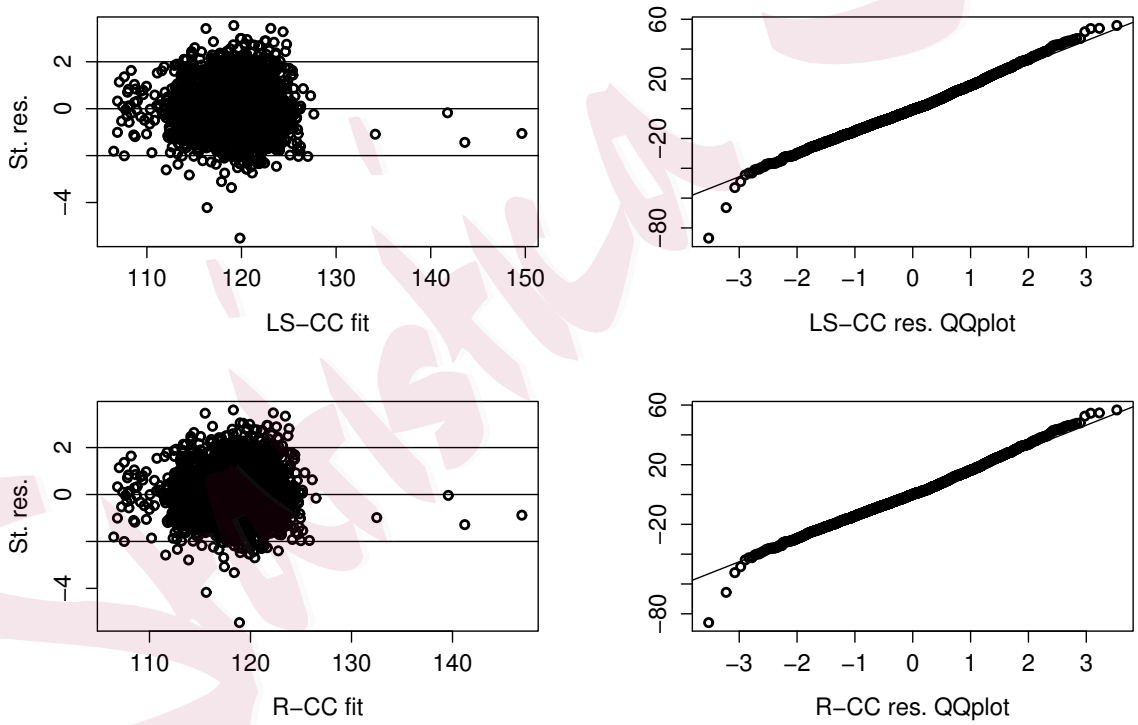

Figure 1: Studentized Residuals plots and Residuals Q-Q plots of the LS (CC) and Rank (CC)

Bibliography

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

BINDELE, H. F. & ABEBE, A. (2012). Bounded influence nonlinear signed-rank regression. *Canadian Journal of Statistics* **40**, 172–189.

BINDELE, H. F. & ABEBE, A. (2015). Semi-parametric rank regression with missing responses. *Journal of Multivariate Analysis* **142**, 117 – 132.

BINDELE, H. F. & ZHAO, Y. (2016). Signed-rank regression inference via empirical likelihood. *Journal of Statistical Computation and Simulation* **86**, 729–739.

BINDELE, H. F. A. (2017). Strong consistency of the general rank estimator. *Communications in Statistics - Theory and Methods,* **46**, 532–539.

BRUNNER, E. & DENKER, M. (1994). Rank statistics under dependent observations and applications to factorial designs. *Journal of Statistical Planning and Inference* **42**, 353 – 378.

CHAMBERS, R. L. & WELSH, A. H. (1993). Log-linear models for survey data with non-ignorable non-response. *Journal of the Royal Statistical Society. Series B (Methodological)* **55**, 157–170.

CHEN, S. X., PENG, L. & QIN, Y.-L. (2009). Effects of data dimension on empirical likelihood. *Biometrika* **96**, 711–722.

DELECROIX, M., HRISTACHE, M. & PATILEA, V. (2006). On semiparametric estimation in single-index regression. *Journal of Statistical Planning and Inference* **136**, 730 – 769.

DIGGLE, P. & KENWARD, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics* **43**, 49–93.

EINMAHL, U. & MASON, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics* **33**, 1380–1403.

FANG, F., ZHAO, J. & SHAO, J. (2016). Imputation-based adjusted score equations in generalized linear models with nonignorable missing covariate values. *Statistica Sinica.* **In Press**.

GONG, Y., PENG, L. & QI, Y. (2010). Smoothed jackknife empirical likelihood method for ROC curve. *Journal of Multivariate Analysis* **101**, 1520 – 1531.

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

HÁJEK, J. & ŠIDÁK, Z. (1967). *Theory of Rank Tests*. New York: Academic Press.

HETTMANSPERGER, T. P. & MCKEAN, J. W. (2011). *Robust Nonparametric Statistical Methods*, vol. 119 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2nd ed.

HJORT, N. L., MCKEAGUE, I. W. & VAN KEILEGOM, I. (2009). Extending the scope of empirical likelihood. *The Annals of Statistics* **37**, 1079–1111.

IBRAHIM, J. G., LIPSITZ, S. R. & CHEN, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**, 173–190.

IBRAHIM, J. G., LIPSITZ, S. R. & HORTON, N. (2001). Using auxiliary data for parameter estimation with non-ignorably missing outcomes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **50**, 361–373.

JAECKEL, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist.* **43**, 1449–1458.

JING, B.-Y., YUAN, J. & ZHOU, W. (2009). Jackknife empirical likelihood. *Journal of the American Statistical Association* **104**, 1224–1232.

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

LAHIRI, S. N., MUKHOPADHYAY, S. et al. (2012). A penalized empirical likelihood method in high dimensions. *The Annals of Statistics* **40**, 2511–2540.

LITTLE, R. J. A. & RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2nd ed.

MA, W.-Q., GENG, Z. & HU, Y.-H. (2003). Identification of graphical models for nonignorable nonresponse of binary outcomes in longitudinal studies. *Journal of Multivariate Analysis* **87**, 24–45.

MIAO, W., DING, P. & GENG, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association.* **In Press**.

MOLENBERGHS, G. & KENWARD, M. (2007). *Missing Data in Clinical Studies*, vol. 61. John Wiley & Sons.

NIU, C., GUO, X., XU, W. & ZHU, L. (2014). Empirical likelihood inference in linear regression with nonignorable missing response. *Computational Statistics & Data Analysis* **79**, 91–112.

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. *The Annals of Statistics* **18**, 90–120.

OWEN, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

QIN, J. & LAWLESS, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* **22**, 300–325.

RAO, B. P. (2009). Conditional independence, conditional mixing and conditional association. *Annals of the Institute of Statistical Mathematics* **61**, 441–460.

ROTNITZKY, A., ROBINS, J. M. & SCHARFSTEIN, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93**, 1321–1339.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

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

TANG, C. Y. & LENG, C. (2010). Penalized high-dimensional empirical likelihood. *Biometrika* **97**, 905–920.

TANG, N., ZHAO, P., QU, A. & JIANG, D. (2016). Semiparametric estimating equations inference with nonignorable nonresponse. *Statistica Sinica.* **In Press**.

TANG, N., ZHAO, P. & ZHU, H. (2014). Empirical likelihood for estimating equations with nonignorably missing data. *Statistica Sinica* **24**, 723.

WANG, Q., LINTON, O. & HÄRDLE, W. (2004). Semiparametric regression analysis with missing response at random. *Journal of the American Statistical Association* **99**, 334–345.

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

WIED, D. & WEIßBACH, R. (2012). Consistency of the kernel density estimator: a survey. *Statistical Papers* **53**, 1–21.

YANG, H. & ZHAO, Y. (2012). New empirical likelihood inference for linear transformation models. *Journal of Statistical Planning and Inference* **142**, 1659–1668.

YANG, H. & ZHAO, Y. (2013). Smoothed jackknife empirical likelihood inference for the difference of ROC curves. *Journal of Multivariate Analysis* **115**, 270–284.

YANG, H. & ZHAO, Y. (2015). Smoothed jackknife empirical likelihood inference for ROC curves with missing data. *Journal of Multivariate Analysis* **140**, 123–138.

YU, W., SUN, Y. & ZHENG, M. (2011). Empirical likelihood method for linear transformation models. *Annals of the Institute of Statistical Mathematics* **63**, 331–346.

ZHANG, Z. & ZHAO, Y. (2013). Empirical likelihood for linear transformation models with interval-censored failure time data. *Journal of Multivariate Analysis* **116**, 398–409.

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