Improved $k$th power expectile regression with nonignorable dropouts

Dongyu Li and Lei Wang

School of Statistics and Data Science & LPMC, Nankai University, Tianjin, People’s Republic of China

ABSTRACT
The $k$th ($1 < k \leq 2$) power expectile regression (ER) can balance robustness and effectiveness between the ordinary quantile regression and ER simultaneously. Motivated by a longitudinal ACTG 193A data with nonignorable dropouts, we propose a two-stage estimation procedure and statistical inference methods based on the $k$th power ER and empirical likelihood to accommodate both the within-subject correlations and nonignorable dropouts. Firstly, we construct the bias-corrected generalized estimating equations by combining the $k$th power ER and inverse probability weighting approaches. Subsequently, the generalized method of moments is utilized to estimate the parameters in the nonignorable dropout propensity based on sufficient instrumental estimating equations. Secondly, in order to incorporate the within-subject correlations under an informative working correlation structure, we borrow the idea of quadratic inference function to obtain the improved empirical likelihood procedures. The asymptotic properties of the corresponding estimators and their confidence regions are derived. The finite-sample performance of the proposed estimators is studied through simulation and an application to the ACTG 193A data is also presented.

1. Introduction
In population health, biological researches, economics, social sciences, data are often collected from every sampled subject at many time points, which are referred to as longitudinal data. Our study is motivated by the following longitudinal data from the AIDS Clinical Trial Group 193A (ACTG 193A), which was a study of HIV-AIDS patients with advanced immune suppression. After the treatment was applied, the CD4 counts were scheduled to be collected from each patient in every 8 weeks. However, because of adverse events, low-grade toxic reactions, the desire to seek other therapies, death and some other reasons, the dropout rates of the first four follow-up times are 31.5%, 42.4%, 55.4% and 65.3%, respectively. In addition, previous experiences from doctors and Cho et al. [3] indicated that a steep decline in the CD4 cell counts indicates the disease progression, and
patients with low CD4 cell counts are more likely to drop out from the scheduled study visits compared to patients with normal CD4 cell counts. Therefore, nonresponse of the CD4 cell counts is likely related to itself and is nonignorable [31]. On the other hand, it can be checked that the distribution of CD4 cell counts is skewed. Hence, the response mean regression may not appropriately assess the longitudinal change in the CD4 cell counts.

Compared to conditional response mean regression, in some studies about the weights in children growth, high expenses in medical cost and so on, we usually postulate the quantile regression (QR; [11]) and expectile regression (ER; [1,18]), which have many advantages over traditional mean regression methods, especially in the presence of heavy-tailed errors. Moreover, both the QR and ER can capture a complete picture of the relationship between the response and predictors. On the other hand, although the least absolute deviation makes the QR method robust, its check function is nondifferentiability at some parameter points which makes computation difficult. Unlike the QR method, the ER replaces the $L_1$ loss by a quadratic $L_2$ loss, which leads to the asymmetric least squares, such that it is easy to compute the estimates but more sensitive to the tails of distributions. Overall, the computation of ER is much easier than that of QR, but the QR is more robust than the ER. In order to simplify computation and balance robustness and effectiveness, the $k$th power ER is proposed by Jiang et al. [9]. In specific, the $k$th power expectile of a random variable $Y$ is defined as the solution $\mu(Y)$ which minimize the loss function $E(Q(Y - \theta))$ over $\theta \in \mathbb{R}$ for a fixed $\tau \in (0,1)$, where

$$Q(u) = |u|^k|\tau - I(u < 0)|,$$

is called as the asymmetric $k$th power loss function. It can be seen that the loss functions of the QR and ER are the two special cases of the $k$th power loss function with $k = 1$ and $k = 2$, respectively. For $1 < k < 2$, the loss function of the $k$th power ER falls in between the loss functions of the QR and ER, which can be termed as a trade-off between the QR and ER methods.

For the longitudinal ACTG 193A data, a major aspect is how to take into account the within-subject correlation structure to improve estimation efficiency. The most commonly used approach is generalized estimating equations (GEEs; Liang and Zeger [13]) based on a working correlation matrix. Recently, Huang et al. [8], Bai et al. [2], Leng et al. [12], Zhang and Leng [32], Fu and Wang [6], Zhang et al. [33], Lv et al. [15] and Xu et al. [30] proposed many methods to approximate or incorporate the within-cluster correlation structure. Among these methods, the quadratic inference function (QIF; Qu et al. [21]) has recently received considerable attention due to its flexibility and simplicity [3,4,27]. Moreover, noticing that the CD4 cell counts drop out of the study prior to the end of the study. The majority of existing methods only naturally accommodate ignorable dropout or missing at random (MAR; [14]) assumption. Unfortunately, developing valid methodologies for statistical analysis with nonignorable dropout is challenging [28]. In this case, the population parameters are, in general, not identifiable [5]) if there is no assumption imposed, and estimators based on the assumption of ignorable dropouts may have large biases. Thus, methods which are very different from those for an ignorable propensity have to be applied; for example, see Molenberghs [17], Kim and Yu [10], Wang et al. [29], Shao and Wang [25] and so on. To address the identifiability issue, Tang et al. [26], Wang et al. [29], Shao and Wang [25] and Wang et al. [28] found that two key assumptions are needed. The first key assumption is that either $p(r_i|x_i,y_i)$ or $f(y_i|x_i)$ has a parametric form [23,28]. The second
key assumption imposed by Tang et al. [26], Wang et al. [29], Shao and Wang [25], Fang and Shao [5], Miao et al. [16] is that \( x_i \) can be decomposed as two parts \( x_i = (u_i, z_i) \) and \( z_i \) is unrelated to dropout propensity conditioned on \( (u_i, y_i) \), i.e. \( \Pr(r_i|x_i, y_i) = \Pr(r_i|u_i, y_i) \). This assumption shows that \( z_i \) introduces additional conditional independence and affects the propensity only through its association with the outcome \( y_i \). More importantly, \( z_i \) can be used to create more instrument-based estimating equations, which helps to identify and estimate the unknown parameters in the propensity [28,29]).

In this paper, we propose a two-stage procedure to construct a class of improved \( k \)th power ER estimators, which can incorporate the within-subject correlations under an informative working correlation structure and account for nonignorable dropouts, by using the bias-corrected empirical likelihood (BEL) in conjunction with the QIF method. In the first stage, we impose a parametric model on the dropout propensity and then use a nonresponse instrument to deal with the identifiability issue. After the dropout propensity is consistently estimated, we construct the bias-corrected GEEs based on the inverse propensity weighting (IPW; [24]). In the second stage, we propose a classes of \( k \)th power ER estimators based on the QIF method, which can incorporate the within-subject correlations under informative working correlation structures.

The resulting EL ratio follows a weighted sum of chi-square random variables asymptotically, which can be used to construct the corresponding confidence regions. Our simulation results show that the proposed estimators are more efficient than estimators obtained under a working independence correlation structure, and the performance of the \( k \)th power ER is better than the performance of the ordinary ER in some data cases.

The rest of this paper is organized as follows. In Section 2, we first present the parametric dropout propensity and instrument approach, and then construct the proposed \( k \)th power ER estimators. The statistical properties are investigated in Section 3. Simulation studies are given in Section 4. Section 5 analyzes the AIDS Clinical Trial Group 193A data for illustration. Some discussions can be found in Section 6. All technical details are provided in the Appendix.

2. Methodology

Let \( y_i = (y_{i1}, y_{i2}, \ldots, y_{imi})^T \) be a \( m_i \) dimensional vector of the \( i \)th subject’s response, which may drop out, \( x_i = (x_{i1}, \ldots, x_{imi})^T \) be an always observed \((m_i \times p)\)-dimensional matrix of covariates associated with \( y_i \), \( r_i = (r_{i1}, r_{i2}, \ldots, r_{imi})^T \) be the vector of response indicators, where \( r_{ij} = 1 \) if \( y_{ij} \) is observed and \( r_{ij} = 0 \) if \( y_{ij}, \ldots, y_{imi} \) are not observed, \( i = 1, \ldots, n \) and \( j = 1, \ldots, m_i \). Here, \( m_i \) is called as the size of the \( i \)th subject. We first investigate the balanced data with the same subject size \( m_i = m \) and the unbalanced situation will be discussed in Section 6 later.

Consider

\[
y_{ij} = x_{ij}^T \beta + \epsilon_{ij},
\]

where \( y_{ij} \) is the \( j \)th measurement on the \( i \)th subject, \( x_{ij} \) is the \( p \times 1 \) covariates, \( \beta \) is \( p \)-vector unknown regression coefficients and \( \epsilon_{ij} \) is the random error for \( i = 1, \ldots, n \) and \( j = 1, \ldots, m \). For a fixed \( \tau \in (0, 1) \), we define the \( k \)th power expectile regression (ER)
where \( \mu(\epsilon_{ij}) = 0 \) is to make sure the error is centered on the \( k \)th expetile. In order to take the nonignorable dropouts into account, define \( \pi_{ij} = \text{Pr}(r_{ij} = 1|x_i, y_i) \). Based on the independence working model assumption and popular IPW approach, an estimator of \( \beta \) can be obtained by minimizing the following bias-corrected objective function

\[
\mathcal{G}(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{r_{ij}}{\pi_{ij}} Q(y_{ij} - x_{ij}^T \beta),
\]

due to the fact \( E\{\frac{r_{ij}}{\pi_{ij}} Q(y_{ij} - x_{ij}^T \beta)\} = E\{Q(y_{ij} - x_{ij}^T \beta)\} \). Motivated by Liang and Zeger [13], the optimal bias-corrected GEEs can be written as

\[
\sum_{i=1}^{n} x_i^T V_i^{-1} W_i S(y_i - x_i \beta) = 0,
\]

where \( W_i = \text{diag}(r_{i1}/\pi_{i1}, \ldots, r_{im}/\pi_{im}) \), \( V_i \) is the covariance matrix of \( S(y_i - x_i \beta) \) and

\[
S(y_i - x_i \beta) = k \begin{pmatrix}
(1)^{1-I(\gamma_{i1} < x_{i1}^T \beta)} \tau - I(\gamma_{i1} < x_{i1}^T \beta) | \gamma_{i1} - x_{i1}^T \beta|^{k-1} \\
\vdots \\
(1)^{1-I(\gamma_{im} < x_{im}^T \beta)} \tau - I(\gamma_{im} < x_{im}^T \beta) | \gamma_{im} - x_{im}^T \beta|^{k-1}
\end{pmatrix}.
\]

Here, \( V_i^{-1} \) can be decomposed as \( A_{i}^{-1/2}(\phi_0)R^{-1}A_i^{-1/2}(\phi_0) \) with \( R \) and \( A_i(\phi_0) \) being the true correlation matrix and marginal variance matrix of \( S(y_i - x_i \beta) \), respectively. \( \phi_0 \) is a true dispersion parameter. Since \( R^{-1} \) is unknown in practice, we borrow the matrix expansion idea of QIF[21] to approximate \( R^{-1} \). In specific, Qu et al. [21] assumed that \( R^{-1} \) can be approximated by a linear combination of several basis matrices, i.e.

\[
R^{-1} = \sum_{j=1}^{q} b_j B_j,
\]

where \( B_1, \ldots, B_q \) are \((m \times m)\)-dimensional symmetric basic matrices depending on the particular choice of \( R^{-1} \) and \( b_1, \ldots, b_q \) are unknown coefficients. More details can be found in Qu et al. [21] and Cho and Qu [4].

To obtain the estimator of \( \beta \), we first need to get consistent estimators of \( \pi_{ij} \) and \( A_i \), denoted as \( \hat{\pi}_{ij} \) and \( \hat{A}_i \), respectively. Using the identifiability approach proposed in Wang et al. [28], we assume that \( x_i \) can be decomposed as \( u_i \) and \( z_i \), i.e. \( x_i = (u_i, z_i) \). Denote \( \tilde{u}_{ij} = (u_{i1}^T, \ldots, u_{ij}^T)^T \), \( \tilde{z}_{ij} = (z_{i1}^T, \ldots, z_{ij}^T)^T \) and \( \tilde{y}_{ij} = (y_{i1}, \ldots, y_{ij})^T \) as the history \( u_{ij}, z_{ij} \) and \( y_{ij} \) up to and including cycle \( j \), respectively. We further assume that the dropout propensity has a parametric form,

\[
\text{Pr}(r_{ij} = 1 | r_{i(j-1)} = 1, x_i, y_i) = \text{Pr}(r_{ij} = 1 | r_{i(j-1)} = 1, -u_{ij}, -y_{ij}) = \Psi(a_j + y_j^T \mathcal{O}_{ij}), \quad j = 1, \ldots, m,
\]

where \( \mathcal{O}_{ij} = (\tilde{u}_{ij}^T, \tilde{y}_{ij}^T)^T \), \( a_j \) is unknown parameter, \( y_j \) is a column vector of unknown parameters, \( \Psi \) is a known monotone function defined on \([0, 1]\), \( r_{i0} \) is always defined to be
1. For any \( j = 1, \ldots, m \), to estimate the unknown parameter vector \( \theta_j = (\alpha_j, y_j^T)^T \), define the instrumental estimating equations

\[
s_j(y_i, x_i, r_i, \theta_j) = r_{i(j-1)} \left\{ \frac{r_{ij}}{\Psi(\alpha_j + y_j^T \Omega_{ij})} - 1 \right\} \left( 1, \hat{u}_{ij}^T, \hat{z}_{ij}^T, \hat{y}_{i(j-1)}^T \right)^T,
\]

and then the two-step generalized method of moments (GMM; Hansen [7]) can be used to obtain the corresponding estimator \( \hat{\theta}_j \). Under model (2), we have

\[
\pi_{ij} = \prod_{t=1}^j \Pr\{r_{it} = 1 | r_{i(t-1)} = 1, \hat{u}_{it}, \hat{y}_{it}\} = \prod_{t=1}^j \Psi(\hat{\alpha}_t + \hat{y}_t^T \Omega_{it})
\]

which can be estimated by \( \hat{\pi}_{ij} = \prod_{t=1}^j \Psi(\hat{\alpha}_t + \hat{y}_t^T \Omega_{it}) \). It can be verified that \( \pi_{ij} \) is a monotone function and continuous almost surely. Hence, by continuous mapping theorem, \( \hat{\pi}_{ij} \) is a consistent estimator of \( \pi_{ij} \) since \( \hat{\theta}_1, \ldots, \hat{\theta}_j \) are consistent estimators of \( \theta_1, \ldots, \theta_j \), respectively [28,29]. Subsequently, a consistent estimator \( \hat{A}_i = A_i(\hat{\phi}) \) can be obtained by computing the sample marginal variance based on the IPW approach with an initial estimate \( \hat{\beta} \) based on (3).

Thus, Equation (4) can be approximated by a linear combination of elements, \( \hat{g}_i(\beta) \), for \( i = 1, \ldots, n \), where

\[
\hat{g}_i(\beta) = \begin{pmatrix}
  x_i^T \hat{A}_i^{-1/2} B_i \hat{A}_i^{-1/2} \hat{W}_i S(y_i - x_i \beta) \\
  \vdots \\
  x_i^T \hat{A}_i^{-1/2} B_q \hat{A}_i^{-1/2} \hat{W}_i S(y_i - x_i \beta)
\end{pmatrix}.
\]

With more estimating equations than the parameters, we propose to apply the following empirical likelihood for the inference of \( \beta \) under some regular conditions. Let \( p_i \) represent the probability weight allocated to \( \hat{g}_i(\beta) \), \( i = 1, \ldots, n \). The empirical log-likelihood ratio function for \( \beta \) is defined as

\[
\hat{R}(\beta) = -2 \sup \left\{ \sum_{i=1}^n \log(n p_i) : p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i \hat{g}_i(\beta) = 0 \right\},
\]

and the maximum EL estimator \( \hat{\beta} \), can be obtained as below:

\[
\hat{\beta} = \arg \min_{\beta} \{ \hat{R}(\beta) \}.
\]

3. Asymptotic theories

**Theorem 3.1:** Under assumptions (C1)–(C4) in the Appendix, as \( n \to \infty \), we have

\[
\sqrt{n}(\hat{\beta} - \beta_0) \rightarrow N(0, V),
\]

where \( V = \{T^T B^{-1} T\}^{-1} T^T B^{-1} C B^{-1} T \{T^T B^{-1} T\}^{-1} \) and \( T, B, C \) are given in the Appendix.
Next, we will study the asymptotic properties of \( \hat{R}(\beta_0) \). Compared to the standard empirical log-likelihood ratio without missing data, the main difference is that the \( \hat{g}_i(\beta_0) \), \( i = 1, \ldots, n \), are not independent and identically distributed. Hence, the asymptotic distribution of \( \hat{R}(\beta_0) \) may not be a standard chi-square. Actually, we will show that \( \hat{R}(\beta_0) \) is asymptotically weighted sum chi-squares.

**Theorem 3.2:** Under the assumptions in Theorem 3.1, as \( n \to \infty \), we have

\[
\hat{R}(\beta_0) \longrightarrow \rho_1 w_1 + \rho_2 w_2 + \ldots + \rho_{p \times q} w_{p \times q},
\]

where \( w_i, i = 1, \ldots, p \times q \) are independent and follow the standard \( \chi^2 \) distribution with one degree, the weights \( \rho_i \) are eigenvalues of \( \mathbf{B}^{-1} \mathbf{C} \).

**Corollary 3.1:** Under the assumptions in Theorem 3.1, let \( r(\beta_0) = (p \times q)/\text{tr}(\mathbf{B}^{-1} \mathbf{C}) \) be an adjustment factor. Along the lines of Rao and Scott [22], it is straightforward to show that \( r(\beta_0)(\rho_1 w_1 + \ldots + \rho_{p \times q} w_{p \times q}) \) is approximately \( \chi^2_{p \times q} \) distributed, i.e.

\[
\hat{R}(\beta_0)r(\beta_0) \longrightarrow \chi^2_{p \times q}.
\]

In general, \( r(\beta) \) is a function of \( \beta \) and the quantity \( r(\beta_0) \) is a measure of information loss due to missing data. An exception occurs when there is no missing data and then \( r(\beta_0) = 1 \).

**4. Simulations**

We consider

\[
y_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + (1 + \gamma |x_{ij2}|) \varepsilon_{ij}, \tag{8}
\]

where \( x_{ij1} \sim N(1, 1), x_{ij2} \sim N(0, 1) \) and are independent, the random errors \( \varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \varepsilon_{i4})^T \) are generated from four different distributions:

1. **(Case 1):** Homoscedastic normal errors: \( \varepsilon_i \sim N(0, \Sigma) \), where \( \Sigma_{jj'} = 4 \rho^{|j-j'|} \) for \( j, j' = 1, \ldots, 4 \).
2. **(Case 2):** Homoscedastic multivariate t errors: \( \varepsilon_i \sim t_{10}(0, \Sigma) \), where \( \Sigma \) is the same as in Case 1.
3. **(Case 3):** Heteroscedastic normal errors: \( \varepsilon_i \sim N(0, \Sigma) \), where \( \Sigma_{jj'} = 3 \rho^{|j-j'|} \) for \( j, j' = 1, \ldots, 4 \).
4. **(Case 4):** Heteroscedastic multivariate t errors: \( \varepsilon_i \sim t_{10}(0, \Sigma) \), where \( \Sigma \) is the same as in Case 3.

We generate these errors with \( \rho=0.7 \) such that \( \varepsilon_i \) are strongly correlated. Under model (8), when the errors are homoscedastic, the \( \tau \) kth power expectile of \( y_{ij} \) given \( x_{ij} \) is

\[
\eta_{y_{ij}}(\tau | x_{ij}) = \beta_0(\tau) + \beta_1(\tau) x_{ij1} + \beta_2(\tau) x_{ij2},
\]

and we have \( \beta = (\beta_0(\tau), \beta_1, \beta_2) = (\beta_0 + \eta(\tau), \beta_1, \beta_2) \), where \( \eta(\tau) = \arg\min_{m \in \mathbb{R}} E(Q(\varepsilon - m)) \) is the \( \tau \) kth power expectile of \( \varepsilon_i \). When the errors are heteroscedastic, we
have
\[
\beta = (\beta_0 + \eta(\tau), \beta_1, \beta_2 + (-1)^{1-I(x_{ij2}>0)}\gamma \eta(\tau)).
\]

Set the true value \((\beta_0, \beta_1, \beta_2) = (0, 1, 2)\). To apply the proposed method, we use the working propensity model as follows:
\[
\Pr(r_{ij} = 1|r_{i(j-1)} = 1, \hat{u}_{ij}, \hat{y}_{ij}) = 1/(1 + \exp(\alpha_j + \gamma_j^T u_{ij} + \gamma_j y_{ij})).
\]

The missing indicators \(r_i = (r_{i1}, r_{i2}, r_{i3}, r_{i4})^T\) are generated from the following three choices:

(M1): \(\Pr(r_{ij} = 1|r_{i(j-1)} = 1, \hat{u}_{ij}, \hat{y}_{ij}) = 1/(1 + \exp(\alpha_j + \gamma_j x_{ij1}))\) with \((\alpha_j, \gamma_j) = (-0.6 + 0.1(j - 1), -0.8 + 0.2(j - 1))\);
(M2): \(\Pr(r_{ij} = 1|r_{i(j-1)} = 1, \hat{u}_{ij}, \hat{y}_{ij}) = 1/(1 + \exp(\alpha_j + \gamma_j^T x_{ij1} + \gamma_j^2 y_{ij}))\) with \((\alpha_j, \gamma_{j1}, \gamma_{j2}) = (-0.2, -0.8 + 0.2(j - 1), -0.4 + 0.1(j - 1))\);
(M3): \(\Pr(r_{ij} = 1|r_{i(j-1)} = 1, \hat{u}_{ij}, \hat{y}_{ij}) = 1/(1 + \exp(\alpha_j + \gamma_j x_{ij1} + \gamma_j^2 y_{ij}))\) with \((\alpha_j, \gamma_{j1}, \gamma_{j2}) = (-0.6 + 0.1(j - 1), -0.8 + 0.2(j - 1), -0.4 + 0.1(j - 1))\).

It can be seen that \(u_{ij} = x_{ij1}\) and the instrument variable \(z_{ij} = x_{ij2}\). In addition, M1 is an ignorable dropout case and the unconditional dropout percentages for four time points computed are about 22.4%, 43.0%, 60.3%, 74.9%; M2 and M3 are two different nonignorable dropout cases, with the computed percentages about 27.8%, 48.7%, 63.1%, 75.7% and 24.7%, 45.4%, 61.2%, 74.8% respectively. The working model is correct under M1 and M2, but misspecified under M3, so that we could see the robustness of the proposed estimators. We compare the following estimators based on \(\hat{g}_i(\beta)\):

(a) the proposed MNAR estimator based on \(\hat{g}_i(\beta)\) in (7) with nonignorable dropout propensity \(\hat{\pi}_{ij}\) in \(\hat{W}_i\). Here, \(\hat{\pi}_{ij} = \prod_{t=1}^{T} \Psi(\hat{\alpha}_t + \hat{\gamma}_t^T \hat{x}_{it})\) is obtained by the two-step GMM method;
(b) the ignorable assumption (MAR) estimator based on \(\hat{g}_i(\beta)\) in (7) with \(\hat{\pi}_{ij} = \prod_{t=1}^{T} \Psi(\hat{\alpha}_t + \hat{\gamma}_t^T \hat{\pi}_{it})\). Here, the ignorable dropout propensity is imposed by a parametric linear logistic regression and the parameters are obtained by the two-step GMM estimator;
(c) the complete case (CC) estimator based on \(\hat{g}_i(\beta)\) with \(\hat{W}_i = diag\{r_{i1}, \ldots, r_{im}\}\);
(d) the full sample (FULL) estimator based on \(\hat{g}_i(\beta)\) with \(\hat{W}_i = I_m\) when there is no missing data, which is used as a standard.

We also explore three common working correlation choices in MNAR: independence (IND), compound symmetry (CS) and first-order autoregressive (AR(1)). For example, if a working correlation structure is CS, then \(R^{-1} = b_1 B_1 + b_2 B_2\), where \(B_1\) is an identity matrix and \(B_2\) is a symmetric matrix with 0 on the diagonal and 1 elsewhere. If \(R^{-1}\) corresponds to AR(1), \(R^{-1} = b_1 B_1 + b_2 B_2 + b_3 B_3\), where \(B_3\) is a symmetric matrix with 1 in elements \((1, 1)\) and \((m, m)\), and 0 elsewhere. Among the three common working correlation choices, the AR(1) structure best approximates the true correlation structure.
4.1. Absolute bias and mean square error

To evaluate the estimation efficiency of these methods, we compute the simulated absolute bias \( (AB) = \frac{\sum_{s=1}^{S} (|\hat{\beta}_{0s}(\tau) - \beta_{0}(\tau)| + |\hat{\beta}_{1s}(\tau) - \beta_{1}(\tau)| + |\hat{\beta}_{2s}(\tau) - \beta_{2}(\tau)|)}{3S} \) and mean square error \( (MSE) = \frac{\sum_{s=1}^{S} \|\hat{\beta}^{s}(\tau) - \beta^{s}(\tau)\|^{2}}{S} \) based on \( S = 1000 \) replications with \( \tau = 0.25, 0.5, 0.75 \) and \( n = 200, 500 \), where \( \hat{\beta}^{s}(\tau) = (\hat{\beta}_{0s}(\tau), \hat{\beta}_{1s}(\tau), \hat{\beta}_{2s}(\tau)) \) is the \( s \)th estimate. Simulation results at \( \tau = 0.5 \) are presented in Figures 1–4 and other results can be found in the Supplemental Material. In each Figure, the four rows correspond to Case 1, Case 2, Case 3 and Case 4, respectively; the three columns correspond to M1, M2 and M3, respectively. There are a few conclusions can be drawn from these simulation results.

(i) **ABs.** For each \( k \), the proposed QIF estimators based on the \( k \)th power ER have good performance, even when the working dropout propensity model is wrong. All the estimators have negligible biases under M1, but the CC and MAR estimators have large biases under nonignorable missing models M2 and M3. When the errors are heteroscedastic, there are similar results as the homoscedastic errors, which shows the proposed estimators are less sensitivity to the error distributions;

(ii) **MSEs.** Under M2 and M3, the proposed QIF estimators have obviously smaller MSEs compared with CC and MAR estimators, even the CC and MAR estimators do not need to estimate their covariance matrices. Under M1, the CC and MAR estimators have slightly smaller MSEs compared with the proposed QIF estimators. The reason is that these estimators do not need to estimate their covariance matrices so that they have small variances and smaller MSEs. Among the three common correlation structures, it can be seen that the AR(1) estimator always has the smallest MSEs and the IND estimator has the largest MSEs. This indicates that we can obtain a more efficient estimator through assuming an informative working correlation structure that may not be correct, rather than ignoring the unknown correlation structure. In addition, it can be seen that the MSEs decrease as the sample size \( n \) becomes large;

(iii) Among three different \( \tau \), the biases and MSEs when \( \tau = 0.25 \) are slightly larger than the biases and MSEs when \( \tau = 0.5 \) and 0.75, but the proposed \( k \)th power ER estimators still perform better than the CC and MAR estimators.

(iv) Among different \( k \) values, it can be seen that the MSEs of the \( k \)th power ER estimators vary slightly, but the best \( k \) selection should be considered in terms of minimizing the MSEs for different errors. For example, under M2, we notice that \( k = 2 \) (the ER) works best for the normal error, \( k = 1.8 \) works best for the heteroscedastic \( t \) error, \( k = 1.9 \) works best for the heteroscedastic normal error. While, for the \( t \) error, we find \( k = 1.6, 1.1, 1.6 \) are suitable values for \( \tau = 0.25, 0.5, 0.75 \), respectively.

(v) Results for \( \tau = 0.25 \) and 0.75 are similar and will be provided in Figures S1–S8 in the Supplementary Material.

In addition, we replace the error distributions Case 2 and Case 4 by the heavy-tailed multivariate \( t \) errors with degrees of freedom 3. The simulated results are given in Figures S9–S16 in the Supplementary Material and the proposed estimators still perform well. Moreover, we extend the quantile adaptive estimators of Tang et al. [27] under the
nonignorable dropout as follows:

$$\tilde{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \hat{p}_i(\hat{\beta}) \sum_{j=1}^{m} \frac{r_{ij}}{\hat{q}_{ij}} Q_\tau (y_{ij} - x_{ij}^T \beta),$$
where $\hat{p}_i(\hat{\beta})$ is firstly obtained based on the proposed $\hat{g}_i(\beta)$ in (7). The simulated MSE ratio of our proposed estimator $\hat{\beta}$ to the estimator $\tilde{\beta}$ by Tang et al. [27], i.e. $MSE(\hat{\beta})/MSE(\tilde{\beta})$, are reported in Tables S1–S2 based on 1000 replications. It can be seen that the MSEs of $\hat{\beta}$ and $\tilde{\beta}$ are quite close. However, the computation of the estimator $\tilde{\beta}$ proposed by Tang et al. [27] is more complex due to its two-step procedure.
Figure 3. ABs of the CC, MAR, MNAR\textsubscript{IND}, MNAR\textsubscript{CS}, MNAR\textsubscript{AR(1)} and FULL estimators when \( \tau = 0.5 \) and \( n = 500 \).

4.2. Coverage probabilities

Under M2 with \( \tau = 0.5 \), we further examine the confidence regions of the kth power ER in terms of the coverage probability (CP) with the best selection k. To construct the confidence
region of $\beta$, we propose to obtain the estimators of $C$ and $B$ by the plug-in method. Based on Corollary 1, the 100$(1 - \alpha)$% confidence region is

$$\text{CI}(\alpha) = \{\beta : r(\hat{\beta})\hat{R}(\beta) < \chi^2_{p \times q, 1 - \alpha}\},$$
where $r(\hat{\beta}) = (p \times q)/\text{tr}(\hat{B}^{-1}\hat{C})$. While, the EL confidence region based on the estimator (b) is obtained similarly by $\text{CI}(\alpha)$ with the ignorable dropout propensity, and the EL confidence regions based on the estimators (c)–(d) are obtained by $\text{CI}(\alpha) = \{\beta : \hat{R}(\beta) < \chi^2_{1-\alpha(p)}\}$ with $\hat{g}_i(\beta)$ in (7) replaced by the corresponding estimating equations under the CC and FULL methods, respectively. According to Qin and Lawless [20], only the full sample method can produce correct EL confidence regions.

The simulated CPs are presented in Table 1 and a few conclusions can be drawn. The coverage probabilities based on the proposed estimators are close to the nominal level 0.99, and are quite comparable to the FULL method assuming no missing data. It can be seen that the coverage probabilities based on the MAR and CC methods have undercoverage.

### 5. Application to HIV-CD4 data

In this section, the proposed method is applied to the ACTG 193A data. The study collected the CD4 counts in every 8 weeks, from 316 patients taken the daily regimen containing 600 mg of zidovudine plus 2.25 mg of zalcitabine. The data set can be obtained from\[http://www.hsph.harvard.edu/fitzmaur/ala/cd4.txt.\] We consider the first four follow-up times, 8, 16, 24, 32, as four time points $j = 1, 2, 3, 4$, and use the CD4 counts in four time intervals, $(4, 12], (12, 20], (20, 28], (28, 36]$, as the study variable $y_{ij}$. In some time intervals, the $i$th patient may have more than one measurement, so we choose the last record as $y_{ij}$ at time point $j$. In addition, we have two continuous covariates, follow up time and age. It is reasonable to treat the follow-up times as $u_{ij}$, which may affect the dropout, and use the baseline covariates ages as instruments $z_{ij}$. The working propensity model (5) with $\Psi(\cdot) = [1 + \exp(\cdot)]^{-1}$ is used. Moreover, we choose the suitable values of $k$ when the $k$th ER estimators have the smallest SEs, and these estimates and their SEs are reported in Table 2.

From Table 2, we find that all the estimates of $\beta_1$ are negative. This is reasonable because the number of patients’ CD4 counts keeps decreasing as time goes on. At the same time, it can be seen that all the estimates of $\beta_2$ are positive. The reason is that these patients’ ages are closely related to survival times and the patients with earlier ages are more likely to drop out of the study. In addition, we find the estimates are close for these three working correlation structures, while CC and MAR estimates have large differences when $\tau = 0.75$.
Table 2. The estimates (with SE in parentheses) based on the CC, MAR and MNAR methods with three working correlation structures for the HIV-CD4 data at $\tau = 0.25$, 0.5, 0.75.

| Methods   | $\beta_0$       | $\beta_1$       | $\beta_2$       |
|-----------|----------------|----------------|----------------|
|           | $\tau = 0.25$  | $\tau = 0.5$   | $\tau = 0.75$  |
| CC        | $-1.789 (0.917)$ | $-1.890 (0.696)$ | $-1.599 (0.540)$ |
| MAR       | $-0.218 (0.198)$ | $-0.326 (0.225)$ | $-0.183 (0.125)$ |
| MRAR(e)   | $1.303 (0.315)$  | $1.723 (0.278)$  | $1.386 (0.198)$  |
| MNARCS    | $-1.708 (0.609)$ | $-0.213 (0.293)$ | $-0.160 (0.297)$ |
| MNARAR(1) | $1.308 (0.262)$  | $1.723 (0.278)$  | $1.360 (0.318)$  |

6. Discussion

The above methods are presented with balanced data, that is, $m_i = m$. For the unbalanced data, the proposed methods in Section 2 can be implemented directly, after using the transformation matrix of Zhou and Qu (2012) to each subject. In addition, we can show that the asymptotic results of Theorems 3.1 and 3.2 still hold for unequal subject sizes using the similar way in the proofs of the theorems.

To apply our proposed methods, finding an instrument is important. In practice, an instrument may be obtained by one of the following three methods. (1) When $y_{ij}$ is obtained under some treatments, baseline covariates measured prior to treatments, such as age group, gender, race, and education level, are related to the study variable but are unrelated with the propensity, when the study variable and other covariates are conditioned, and thus can be considered as instruments. (2) A fully observed proxy or a mismeasured version of the missing response may be treated as an instrument. (3) A sensitivity analysis of studying some or all different decompositions of $x_i = (u_i, z_i)$ [25].

Turning to the optimal $k$ selection, we can calculate $\hat{V}$ in Theorem 3.1 for each value of $k$ by the plug-in method and then find a favorable $k$ value, for example $k_0$, such that the sum of diagonal elements of $\hat{V}$ of the corresponding $k_0$th power expectilie regression is smallest. We may adopt a simple bootstrap approach by resampling the paired observations $(x_i, y_i, \delta_i), i = 1, \ldots, n$, with replacement to approximate the asymptotic variance. For $b = 1, \ldots, B$, based on the bootstrapped sample $(x_i^b, y_i^b, \delta_i^b), i = 1, \ldots, n$, we obtain $\hat{\beta}_b$ through minimizing

$$g^*_i(\beta) = \left( x_i^T (\hat{A}_i^b)^{-1/2} B_1 (\hat{A}_i^b)^{-1/2} \hat{W}_i^* S (y_i^b - x_i^b \beta) \right)$$

$$\vdots$$

$$\left( x_i^T (\hat{A}_i^b)^{-1/2} B_q (\hat{A}_i^b)^{-1/2} \hat{W}_i^* S (y_i^b - x_i^b \beta) \right)$$
\[ \hat{\beta} = B^{-1} \sum_{b=1}^{B} \hat{\beta}_b \]
and its variance–covariance estimator
\[ \hat{V} = (B(B-1))^{-1} \sum_{b=1}^{B} (\hat{\beta}_b - \hat{\beta})(\hat{\beta}_b - \hat{\beta})^T. \]
Alternatively, based on our simulation results, if we have Gaussian-like distribution of errors, choose \( 1.5 \leq k \leq 1.9 \); if we have heavy-tailed errors, choose \( 1.1 \leq k \leq 1.5 \); if we have nothing about the errors, \( k = 1.5 \) is appealing as a compromise between the robustness of \( k = 1 \) and the efficiency of \( k = 2 \).

**Acknowledgments**
We are grateful to the Editor, an associate editor and two anonymous referees for their insightful comments and suggestions, which have led to significant improvements.

**Disclosure statement**
No potential conflict of interest was reported by the author(s).

**Funding**
This paper was supported by the National Natural Science Foundation of China [grant numbers 11871287, 11771144, 11801359], the Natural Science Foundation of Tianjin [18JCYBJC41100], Fundamental Research Funds for the Central Universities and the Key Laboratory for Medical Data Analysis and Statistical Research of Tianjin.

**References**

[1] D.J. Aigner, T. Amemiya, and D.J. Poirier, *On the estimation of production frontiers: Maximum likelihood estimation of the parameters of a discontinuous density function*, Int. Econ. Rev. (Philadelphia) 17 (1976), pp. 377–396.

[2] Y. Bai, W.K. Fung, and Z. Zhu, *Weighted empirical likelihood for generalized linear models with longitudinal data*, J. Statist. Plann. Inference. 140 (2010), pp. 3446–3456.

[3] H. Cho, H.G. Hong, and M.O. Kim, *Efficient quantile marginal regression for longitudinal data with dropouts*, Biostatistics 17 (2016), pp. 561–575.

[4] H. Cho and A. Qu, *Efficient estimation for longitudinal data by combining large-dimensional moment conditions*, Electron. J. Stat. 9 (2015), pp. 1315–1334.

[5] G.M. Fitzmaurice, G. Molenberghs, and S.R. Lipsitz, *Regression models for longitudinal binary responses with informative drop-outs*, J. R. Stat. Soc. 57 (1995), pp. 691–704.

[6] L. Fu and Y. Wang, *Quantile regression for longitudinal data with a working correlation model*, Comput. Statist. Data Anal. 56 (2012), pp. 2526–2538.

[7] L.P. Hansen, *Large sample properties of generalized method of moments estimators*, Econometrica 50 (1982), pp. 1029–1054.

[8] J. Huang, L. Liu, and N. Liu, *Estimation of large covariance matrices of longitudinal data with basis function approximations*, J. Comput. Graph. Stat. 16 (2007), pp. 189–209.

[9] Y. Jiang, F. Lin, and Y. Zhou, *The kth power expectile regression*, Ann. Inst. Statist. Math. 73 (2021), pp. 83–113.

[10] J.K. Kim and C.L. Yu, *A semiparametric estimation of mean functionals with nonignorable missing data*, J. Amer. Statist. Assoc. 106 (2011), pp. 157–165.

[11] R. Koenker and J.G. Bassett, *Regression quantiles*, Econometrica 46 (1978), pp. 33–50.

[12] C. Leng, W. Zhang, and J. Pan, *Semiparametric mean-covariance regression analysis for longitudinal data*, J. Amer. Statist. Assoc. 105 (2010), pp. 181–193.

[13] K. Liang and S. Zeger, *Longitudinal data analysis using generalized linear models*, Biometrika 73 (1986), pp. 13–22.

[14] R.J.A. Little and D.B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed.Wiley, New York, 2002.
Appendix

(C1) The parameter vector \( \beta_0 \) is an interior point of a compact subset of \( \mathbb{R}^p \), \( \beta_0 \in \mathcal{B} \) is the unique solution to \( E g_i(\beta) = 0 \). \( ||\partial g_i(\beta) / \partial \beta|| \), \( ||\partial^2 g_i(\beta) / \partial \beta \partial \beta^\top|| \) and \( ||g_i(\beta)||^3 \) are bounded by an integrable function \( H \) on a neighborhood of \( \beta_0 \), where \( g_i(\beta) = x_i^\top A_i^{-1/2} B_i A_i^{-1/2} W_i S(y_i - x_i \beta) \) is the \( j \)th element of \( g_i(\beta) \).

(C2) The estimated scale parameter \( \hat{\phi} \) is \( \sqrt{n} \)-consistent given \( \beta \), that is, \( n^{1/2}(\hat{\phi} - \phi_0) = O_p(1) \), \( ||\partial g_i(\beta, \phi) / \partial \phi|| \leq G(\gamma, \beta) \) for \( i = 1, \ldots, n \), where \( G(\cdot, \cdot) \) is function of sample \( y \) and \( \beta \) with \( O_p(1) \).

(C3) For \( j = 1, \ldots, m \), let \( \Theta_j = (\theta_1^j, \ldots, \theta_m^j)^\top \). The response probability function \( \pi_{ij}(\Theta_j) \) satisfies (a) it is twice differentiable with respect to \( \Theta_j \); (b) \( 0 < c_0 < \pi_{ij}(\Theta_j) < 1 \) for a positive constant \( c_0 \); (c) \( \partial \pi_{ij}(\Theta_j) / \partial \Theta_j \) is uniformly bounded.
indistributions, where \( \theta \) is the marginal variance matrix of \( x \). Proof of Lemma A.1.

\( \theta = \) the unique solution to \( E[s_j(y_i, x_i, r_i, \theta_i)] = 0 \) and models (1)-(2) hold, \( \Omega_j = E[s_j(y_i, x_i, r_i, \theta_0^j)s_j(y_i, x_i, r_i, \theta_0^j)^T] \) is positive definite and the matrix \( A_j = E[\partial s_j(y_i, x_i, r_i, \theta_0^j)/\partial \theta_j] \) is of full rank. As \( n \to \infty \), \( \sqrt{n} (\hat{\theta}_j - \theta_j) \to N(0, (A_j^T \Omega_j A_j)^{-1}) \) and \( \sqrt{n} (\hat{\Theta} - \Theta) \to N(0, \Sigma) \) in distributions, where \( \theta_j^0 \) is the true value of \( \theta_j \) and \( \hat{\Theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)^T \).

**Lemma A.1:** Under assumptions (C1)–(C4) in the Appendix, we have

\[
\begin{align*}
(1) \quad & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \to N(0, C), \\
(2) \quad & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \hat{g}_i(\beta_0)^T \overset{p}{\to} B, \\
(3) \quad & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \hat{g}_i(\beta_0)}{\partial \beta} \overset{p}{\to} T, \\
(4) \quad & \max_i ||\hat{g}_i(\beta_0)|| = o_p(n^{1/2}),
\end{align*}
\]

where \( F = (A_1^{-1/2} B_1 A_1^{-1/2}, \ldots, A_m^{-1/2} B_m A_m^{-1/2})^T \), \( C = B + \Delta \Sigma \Delta^T, B = E(FHF^T), \Delta = E(FD) \), \( A \) is the marginal variance matrix of \( (y - x\beta_0) \) substituted by population, \( D \) and \( H \) are given in the proof.

**Proof of Lemma A.1.** Denote \( F_i = (A_1^{-1/2}(\phi_0) B_1 A_1^{-1/2}(\phi_0) x_i, \ldots, A_m^{-1/2}(\phi_0) B_m A_m^{-1/2}(\phi_0) x_i)^T \). To prove (1), using the condition (C2) and Taylor expansion, we have

\[
\begin{align*}
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{g}_i(\beta_0) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i(\beta_0, \hat{\Theta}, \hat{\phi}) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i(\beta_0, \hat{\Theta}, \phi_0) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial g_i(\beta_0, \hat{\Theta}, \phi^*)}{\partial \phi} (\hat{\phi} - \phi_0) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i(\beta_0, \hat{\Theta}, \phi_0) + o_p(1),
\end{align*}
\]

where \( \phi^* \) lies between \( \phi_0 \) and \( \hat{\phi} \). Then we obtain

\[
\begin{align*}
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{g}_i(\beta_0) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i(\beta_0, \hat{\Theta}, \phi_0) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} F_i \hat{W}_i S(y_i - x_i \beta_0) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} F_i W_i S(y_i - x_i \beta_0) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} F_i (\hat{W}_i - W_i) S(y_i - x_i \beta_0) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_i + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} T_i^1 + o_p(1) \\
&=: I + I^*.
\end{align*}
\]

For \( I_i \), note that \( E(I_i) = E[F_i W_i S(y_i - x_i \beta_0)] = E[E[F_i W_i S(y_i - x_i \beta_0)|y_i, x_i]] = E[F_i S(y_i - x_i \beta_0)] \), and similarly we have \( E(I_i^1) = E[F_i S(y_i - x_i \beta_0)S^T(y_i - x_i \beta_0)F_i^T] = E[F_i H_i F_i^T] \), where \( H_i \)
is a \((m \times m)\)-dimensional matrix. It can be verified that
\[
I \longrightarrow N(0, B),
\]
with \(B = E[FHF^T]\). Here, \(F\) and \(H\) denote the matrix substituted by population. Denote \(\Theta_j^0\) as the true parameters vector, \(j = 1, 2, \ldots, m\). Similar to the proof of \(E(I_i)\), it leads to
\[
E(I_i^*) = E[F_i(\hat{W}_i - W_i)S(y_i - x_i\beta_0)]
\]
\[
= E\left[ F_i \text{diag} \left( \frac{\partial \pi_i(\Theta_i^0)}{\partial \Theta_i} \right) \right] S(y_i - x_i\beta_0) + o_p(n^{-1/2})
\]
\[
= E\left[ F_iD_i(\hat{\Theta}_m - \Theta^0_m) \right] + o_p(n^{-1/2}),
\]
where
\[
D_i = \begin{pmatrix}
S(y_{i1} - x_{i1}\beta_0) \\
\vdots \\
S(y_{im} - x_{im}\beta_0)
\end{pmatrix}
\]
Invoking with \((\hat{\Theta}_m - \Theta^0_m) = O_p(n^{-1/2})\), we have \(E(I_i^*) = O_p(n^{-1/2})\). As a result,
\[
I^* \longrightarrow N(0, \Delta\Sigma\Delta^T),
\]
where \(\Sigma\) is the asymptotic covariance matrix of \(\sqrt{n}(\hat{\Theta}_m - \Theta_m)\), \(\Delta = E(FD)\) and \(D\) denotes the matrix substituted by population. Further, it can be verified that \(E(I^*n) = o_p(1)\) and \(\text{Cov}(I, I^*) = o_p(1)\), which lead to \(\text{Var}(I + I^*) = B + \Delta\Sigma\Delta^T\). Then it follows that
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \longrightarrow N(0, B + \Delta\Sigma\Delta^T).
\]
To prove (2), we can show that
\[
\frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0)\hat{g}_i(\beta_0)^T = \frac{1}{n} \sum_{i=1}^{n} F_iH_iF_i^T + \frac{1}{n} \sum_{i=1}^{n} F_i(\hat{H}_i - H_i)F_i^T
\]
\[
= J_n + J_n^*.
\]
For the term \(J_n\), we have \(J_n \overset{p}{\longrightarrow} B\) with \(B = E(FHF^T)\). For the term \(J_n^*\), similar to the proof of \(E(I_i^*)\), it can be verified that \(EJ_n^* = O_p(n^{-1/2})\). Thus, it follows
\[
\frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0)\hat{g}_i(\beta_0)^T \overset{p}{\longrightarrow} B.
\]
To prove (3), we proceed \(E\hat{\partial}\hat{g}_i(\beta_0)/\hat{\partial}\beta\) as follows:
\[
E\left( \frac{\partial \hat{g}_i(\beta_0)}{\partial \beta} \right) = E \frac{\partial}{\partial \beta} \left( F_iW_i(S(y_i - x_i\beta_0)) \right) E \frac{\partial}{\partial \beta} \left( F_i(\hat{W}_i - W_i)S(y_i - x_i\beta_0) \right)
\]
\[
= T_i + T_i^*.
\]
For $T_i$, we have
\[
T_i = E \frac{\partial}{\partial \beta} [F_i W_i S(y_i - x_i \beta_0)]
= E[F_i W_i S'(y_i - x_i \beta_0) x_i],
\]
where $S'(y_i - x_i \beta_0) = (k - 1) \text{diag}(|\tau - I(y_{I1} < x_{i1} \beta_0)| |y_{i1} - x_{i1} \beta_0|^{k-2}, \ldots, |\tau - I(y_{im} < x_{im} \beta_0)| |y_{im} - x_{im} \beta_0|^{k-2})$. Similar to the proof of $EI^*_i$, we have $ET_i^* = O_p(n^{-1/2})$. This yields the result that
\[
\frac{1}{n} \sum_{i=1}^n \frac{\partial \hat{g}_i(\beta_0)}{\partial \beta} \xrightarrow{p} T,
\]
where $T$ is the matrix substituted by population. According to (2),
\[
\frac{\left(\max_i \|\hat{g}_i(\beta_0)\|/n\right)^2}{\sum_{i=1}^n \hat{g}_i(\beta_0) \hat{g}_i(\beta_0)^T/n} \rightarrow 0,
\]
which leads to $\max_i \|\hat{g}_i(\beta_0)\| = O_p(n^{1/2})$.

**Lemma A.2:** Under assumptions (C1)–(C4) in the Appendix, then as $n \to \infty$, with probability to 1, $\hat{R}(\beta)$ has a minimum in the interior of the ball $||\beta - \beta_0|| \leq n^{-1/3}$.

**Proof of Lemma A.2.** Let $\beta = \beta_0 + ud_n$, where $\|u\| = 1$. For simplicity, we denote $\lambda = \lambda(\beta)$. First, we give a lower bound for $\hat{R}(\beta)$ on the surface of the ball. Similar to the proof of Owen [19], we have
\[
\hat{R}(\beta) = 2 \sum_{i=1}^n \lambda^T \hat{g}_i(\beta) - \sum_{i=1}^n \lambda^T \hat{g}_i(\beta) \hat{g}_i(\beta)^T \lambda + o(n^{1/3})
\]
\[
= 2n \left\{ \frac{1}{n} \sum_{i=1}^n \hat{g}_i(\beta) \right\}^T \left\{ \frac{1}{n} \sum_{i=1}^n \hat{g}_i(\beta) \hat{g}_i(\beta)^T \right\}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n \hat{g}_i(\beta) \right\} + o(n^{1/3})
\]
\[
= 2n \left\{ \frac{1}{n} \sum_{i=1}^n \hat{g}_i(\beta_0) + \frac{1}{n} \sum_{i=1}^n \frac{\partial \hat{g}_i(\beta_0)}{\partial \beta^T} u n^{-1/3} \right\}^T \left\{ \frac{1}{n} \sum_{i=1}^n \hat{g}_i(\beta_0) \hat{g}_i(\beta_0)^T \right\}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n \hat{g}_i(\beta_0) + \frac{1}{n} \sum_{i=1}^n \frac{\partial \hat{g}_i(\beta_0)}{\partial \beta^T} u n^{-1/3} \right\} + o(n^{1/3})
\]
\[
= 2n \left\{ O(n^{-1/2} (\log \log n)^{1/2}) + E \frac{\partial \hat{g}_i(\beta)}{\partial \beta^T} u n^{-1/3} \right\}^T \left\{ \frac{1}{n} \sum_{i=1}^n \hat{g}_i(\beta) \hat{g}_i(\beta)^T \right\}^{-1} \left\{ O(n^{-1/2} (\log \log n)^{1/2}) + E \frac{\partial \hat{g}_i(\beta)}{\partial \beta^T} u n^{-1/3} \right\} + o(n^{1/3}).
\]
This yields $\hat{R}(\beta) \geq (c - \epsilon)n^{1/3}$, a.s., where $c - \epsilon > 0$ and $c$ is the smallest eigenvalue of $E[\partial \hat{g}_i(\beta_0)/\partial \beta^T]^T (E \hat{g}_i(\beta_0) \hat{g}_i(\beta_0)^T)^{-1} E[\partial \hat{g}_i(\beta_0)/\partial \beta^T]$. Similarly, $\hat{R}(\beta_0) = O(\log \log n)$, a.s. Since $\hat{R}(\beta)$ is a continuous function about $\beta$, $\hat{R}(\beta)$ has a minimum value in the interior of this ball $||\beta - \beta_0|| \leq n^{-1/3}$. ■
**Proof of Theorem 3.1.** Note that the Lagrange multiplier method leads to the smoothed empirical log-likelihood ratio function for \( \beta_0 \)

\[
\hat{R}(\beta_0) = 2\sum_{i=1}^{n} \log(1 + \lambda^T(\beta_0)\hat{g}_i(\beta_0)),
\]

where vector \( \lambda = \lambda(\beta_0) \) is the solution to

\[
h(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{g}_i(\beta_0)}{1 + \lambda^T\hat{g}_i(\beta_0)} = 0.
\]

Under certain conditions, \( \hat{\beta} \) is the solution by simultaneously satisfying the following two equations:

\[
T_{1n}(\beta, \lambda) = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{g}_i(\beta)}{1 + \lambda^T\hat{g}_i(\beta)} = 0,
\]

\[
T_{2n}(\beta, \lambda) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \hat{g}_i(\beta)/\partial \beta}{1 + \lambda^T\hat{g}_i(\beta)} = 0.
\]

Note that \( T_{1n}(\beta_0, 0) = n^{-1}\sum_{i=1}^{n} \hat{g}_i(\beta_0) \) and \( T_{2n}(\beta_0, 0) = 0 \). Applying Taylor expansions to \( T_{1n}(\hat{\beta}, \hat{\lambda}) \) and \( T_{2n}(\hat{\beta}, \hat{\lambda}) \) at \((\beta_0, 0)\), we can get

\[
0 = T_{1n}(\beta_0, 0) + \frac{\partial T_{1n}(\beta_0, 0)}{\partial \beta}(\hat{\beta} - \beta_0) + \frac{\partial T_{1n}(\beta_0, 0)}{\partial \lambda^T}\hat{\lambda} + o_p(u_n),
\]

\[
0 = T_{2n}(\beta_0, 0) + \frac{\partial T_{2n}(\beta_0, 0)}{\partial \beta}(\hat{\beta} - \beta_0) + \frac{\partial T_{2n}(\beta_0, 0)}{\partial \lambda^T}\hat{\lambda} + o_p(u_n),
\]

where \( u_n = ||\hat{\beta} - \beta_0|| + ||\hat{\lambda}|| \). Also, the above two equations can be rewritten as

\[
\begin{pmatrix} \hat{\lambda} \\ \hat{\beta} - \beta_0 \end{pmatrix} = \begin{pmatrix} \frac{\partial T_{1n}(\beta_0, 0)}{\partial \lambda^T} & \frac{\partial T_{1n}(\beta_0, 0)}{\partial \beta} \\ \frac{\partial T_{2n}(\beta_0, 0)}{\partial \lambda^T} & 0 \end{pmatrix}^{-1} \begin{pmatrix} -\frac{1}{n}\sum_{i=1}^{n} \hat{g}_i(\beta_0) + o_p(u_n) \\ \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \hat{g}_i(\beta_0)}{\partial \beta} \end{pmatrix}.
\]

From this and \( T_{1n}(\beta_0, 0) = n^{-1}\sum_{i=1}^{n} \hat{g}_i(\beta_0) = O_p(n^{-1/2}) \), we know that \( u_n = O_p(n^{-1/2}) \). Then we have

\[
\sqrt{n}(\hat{\beta} - \beta_0) = \{T^TB^{-1}T\}^{-1}T^TB^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{g}_i(\beta_0) + o_p(1).
\]

Together with Lemma A.1, we can derive that

\[
\sqrt{n}(\hat{\beta} - \beta_0) \longrightarrow N(0, \{T^TB^{-1}T\}^{-1}T^TB^{-1}CB^{-1}T(T^TB^{-1}T)^{-1}).
\]

**Proof of Theorem 3.2.** By the proof of Lemma A.1, and applying the similar arguments used in the proof of (2.14) in Owen [19], we can show that

\[
||\lambda|| = O_p(n^{-1/2}).
\]
Write $\lambda = \lambda(\beta_0) = \rho u$, where $\rho = ||\lambda||$, $u = \lambda / ||\lambda||$ and $||u|| = 1$. We have

$$0 = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{g}_i(\beta_0)}{1 + \lambda^T (\hat{g}_i(\beta_0))} = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{g}_i(\beta_0)}{1 + \rho u^T \hat{g}_i(\beta_0)} = \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0) - \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \hat{g}_i(\beta_0)^T u \rho.$$

By multiplying $u^T$, it leads to

$$\left\| u^T \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \right\| = \frac{1}{n} \sum_{i=1}^{n} u^T \hat{g}_i(\beta_0) \hat{g}_i(\beta_0)^T u \rho \geq \frac{1}{1 + \rho \max_{1 \leq i \leq n} |\hat{g}_i(\beta_0)|} \frac{1}{n} \sum_{i=1}^{n} u^T \hat{g}_i(\beta_0) \hat{g}_i(\beta_0)^T u \rho,$$

where the inequality follows from positivity of $1 + \rho u^T \hat{g}_i(\beta_0)$. Then

$$\rho \frac{1}{n} \sum_{i=1}^{n} u^T \hat{g}_i(\beta_0) \hat{g}_i(\beta_0)^T u \leq \left\| u^T \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \right\| [1 + \rho \max_{1 \leq i \leq n} |\hat{g}_i(\beta_0)|].$$

Note that $\max_{1 \leq i \leq n} ||\hat{g}_i(\beta_0)|| = o(n^{1/2})$. By WLLN and Lemma A.1, we have

$$\rho [u^T Bu + o_p(1)] \leq \left\| u^T \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \right\| + \rho o(n^{1/2}),$$

which leads to $\rho = O_p(n^{-1/2})$, i.e. $||\lambda|| = O_p(n^{-1/2})$. Naturally,

$$\max_{1 \leq i \leq n} ||\lambda^T \hat{g}_i(\beta_0)|| \leq ||\lambda|| \max_{1 \leq i \leq n} ||\hat{g}_i(\beta_0)|| = O_p(n^{-1/2}) o(n^{1/2}) = o_p(1).$$

Expanding $h(\lambda)$, we get

$$0 = h(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \left\{ 1 - \lambda^T \hat{g}_i(\beta_0) + \frac{[\lambda^T \hat{g}_i(\beta_0)]^2}{(1 + \xi_i)^3} \right\} = \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0) - \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \hat{g}_i(\beta_0)^T \lambda + \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \frac{[\lambda^T \hat{g}_i(\beta_0)]^2}{(1 + \xi_i)^3},$$

where $\xi_i \in (0, \lambda^T \hat{g}_i(\beta_0))$. Using the fact that $\max_{1 \leq i \leq n} ||\lambda^T \hat{g}_i(\beta_0)|| = o_p(1)$, then $|\xi_i| = o_p(1)$. Note that

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0) [\lambda^T \hat{g}_i(\beta_0)]^2 \right\| \leq \frac{\max_{1 \leq i \leq n} ||\hat{g}_i(\beta_0)||}{1 - \max_{1 \leq i \leq n} |\xi_i|} \left\| \lambda^T \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \hat{g}_i(\beta_0)^T \lambda \right\| = o(n^{1/2}) O_p(n^{-1}) = o_p(n^{-1/2}),$$

thus

$$\lambda = \left[ \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \hat{g}_i(\beta_0)^T \right]^{-1} \left[ \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \right] + \xi,$$
where \(|\xi| = o_p(n^{1/2}/h^{2r} + n^{-1/2})\). A Taylor expansion of \(\hat{R}(\beta_0)\) yields

\[
\hat{R}(\beta_0) = 2 \sum_{i=1}^{n} \left\{ \lambda^T \hat{g}_i(\beta_0) - \frac{1}{2} [\lambda^T \hat{g}_i(\beta_0)]^2 + \frac{1}{3} \frac{[\lambda^T \hat{g}_i(\beta_0)]^3}{(1 + \xi_i)^3} \right\}
\]

\[
= 2\lambda^T \sum_{i=1}^{n} \hat{g}_i(\beta_0) - \sum_{i=1}^{n} \lambda^T \hat{g}_i(\beta_0) \hat{g}_i(\beta_0)^T \lambda + \frac{2}{3} \sum_{i=1}^{n} \frac{[\lambda^T \hat{g}_i(\beta_0)]^3}{(1 + \xi_i)^3}.
\]

Similarly,

\[
\left\| \sum_{i=1}^{n} \frac{[\lambda^T \hat{g}_i(\beta_0)]^3}{(1 + \xi_i)^3} \right\| \leq \max_{1 \leq i \leq n} \left\| \frac{[\lambda^T \hat{g}_i(\beta_0)]^3}{(1 + \xi_i)^3} \right\| \lambda^T \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \hat{g}_i(\beta_0)^T \lambda
\]

\[
= o_p(1) \|O_p(n^{-1})
\]

\[
= o_p(1),
\]

therefore,

\[
\hat{R}(\beta_0) = 2\lambda^T \sum_{i=1}^{n} \hat{g}_i(\beta_0) - \sum_{i=1}^{n} \lambda^T \hat{g}_i(\beta_0) \hat{g}_i(\beta_0)^T \lambda + o_p(1).
\]

(A2)

Substituting (A1) in (A2), it holds that

\[
\hat{R}(\beta_0) = n \left[ \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0)^T \right] \left[ \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \hat{g}_i(\beta_0)^T \right]^{-1} \left[ \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \right]
\]

\[- n\xi^T \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \hat{g}_i(\beta_0)^T \xi + o_p(1).
\]

A simple calculation shows that

\[
n\xi^T \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \hat{g}_i(\beta_0)^T \xi = n_o_p(n^{-1/2})o_p(n^{-1/2}) = o_p(1).
\]

Therefore,

\[
\hat{R}(\beta_0) = \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{g}_i(\beta_0)^T \right] \left[ \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \hat{g}_i(\beta_0)^T \right]^{-1} \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{g}_i(\beta_0) \right] + o_p(1).
\]

Together with the proof of Lemma A.1, we can conclude that

\[
\hat{R}(\beta_0) \rightarrow \rho_1 w_1 + \rho_2 w_2 + \ldots + \rho_{p \times q} w_{p \times q},
\]

where \(w_i, i = 1, \ldots, p \times q\) are independent and follow the standard \(\chi^2\) distribution with one degree, the weights \(\rho_i\) are eigenvalues of \(B^{-1}C\).

\[\blacksquare\]