*Biostatistics* (2016), **17**, 3, *pp.* 561–575
doi:10.1093/biostatistics/kxw007
Advance Access publication on March 7, 2016

# Efficient quantile marginal regression for longitudinal data with dropouts

HYUNKEUN CHO

*Department of Statistics, Western Michigan University, Kalamazoo, MI 49008, USA*

HYOKYOUNG GRACE HONG

*Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA*

MI-OK KIM*

*Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA*

miok.kim@cchmc.org



## Summary

In many biomedical studies independent variables may affect the conditional distribution of the response differently in the middle as opposed to the upper or lower tail. Quantile regression evaluates diverse covariate effects on the conditional distribution of the response with quantile-specific regression coefficients. In this paper, we develop an empirical likelihood inference procedure for longitudinal data that accommodates both the within-subject correlations and informative dropouts under missing at random mechanisms. We borrow the matrix expansion idea of the quadratic inference function and incorporate the within-subject correlations under an informative working correlation structure. The proposed procedure does not assume the exact knowledge of the true correlation structure nor does it estimate the parameters of the correlation structure. Theoretical results show that the resulting estimator is asymptotically normal and more efficient than one attained under a working independence correlation structure. We expand the proposed approach to account for informative dropouts under missing at random mechanisms. The methodology is illustrated by empirical studies and a real-life example of HIV data analysis.

*Keywords*: Empirical likelihood; Longitudinal data; Missing at random; Quadratic inference function; Quantile regression.

## 1. Introduction

Longitudinal data arise frequently in epidemiology, medical science and socioeconomic panel studies, where repeated measurements within the same subject are likely to be correlated. The Gaussian paradigm dominates the analysis of longitudinal data, whereas in many cases the correlated responses follow a non-normal distribution. Moreover, the assumption that the independent variables uniformly affect the different

*To whom correspondence should be addressed.

parts of the conditional distribution of the response may make little sense in biomedical application. For example, in a clinical trial of HIV disease which evaluates effects of different treatments on CD4 cell counts longitudinally, the treatment effects on study subjects with high CD4 cell counts may differ from the treatment effects on study subjects with low CD4 counts who are much sicker. As exhibited in the motivating example of this paper, the distribution of CD4 cell counts is also highly skewed.

Quantile regression provides a viable alternative; it estimates diverse effects of independent variables with quantile-specific regression coefficients without imposing any distributional assumption on the responses. Recent developments in quantile regression approaches for longitudinal data include a quasi-likelihood approach to median regression (Jung, 1996), Bayesian modeling (Dunson *and others*, 2003), a penalized least squares approach (Koenker, 2004), inference via a random intercept through the asymmetric Laplace density (Geraci and Bottai, 2007), weighted quantile regression using a stationary autocorrelation structure (Lu and Fan, 2015), and references therein. Most of the work, however, falls short from readily accommodating two common features of the longitudinal data, correlations between repeated measurements and dropouts.

Unbalanced longitudinal data are quite common due to dropouts, and observed data often provide information on them. In the motivating HIV study example, a proportion of the participants were lost to follow-up. As Volberding *and others* (1990) suggested, it might be due to selective withdrawal of patients with low or declining CD4 cell counts. Missing data caused by such dropouts are missing at random and are challenging with quantile regression. The majority of existing methods take the framework of generalized estimating equations (Liang and Zeger, 1986), which only naturally accommodates missing data completely at random. Alternatively Lipsitz *and others* (1997) and Robins *and others* (1995) employed a weighted generalized estimating equation for monotone missing data assuming that measurements within the same subject are independent.

In this paper, we develop an empirical likelihood-based inference procedure for the marginal quantile regression which accommodates both the within-subject correlations and dropouts under missing at random mechanisms. We use the matrix expansion idea of quadratic inference function (Qu *and others*, 2000) and construct constraints of the empirical likelihood procedure that incorporates the within-subject correlations under an informative working correlation structure. This feature contrasts with those of existing empirical likelihood approaches to the marginal quantile regression, most of which takes the framework of generalized estimating equations under a working independence correlation and forgoes the opportunity of utilizing the within-subject correlations (e.g. Wang and Zhu, 2011; Whang, 2006). We further expand the procedure in order to account for dropouts that are missing at random. We model the dropout process and incorporate the missing data information as weights for the constraints of the empirical likelihood. With the weighted constraints hence defined, the proposed empirical likelihood inference procedure is readily implemented by existing R packages *emplik* and *optim*.

If missing data are not considered, Tang and Leng (2011) also used the empirical likelihood and the matrix expansion ideas of the quadratic inference function, and proposed a two-step approach to the quantile marginal regression. In the first step they constructed the empirical likelihood for the conditional mean regression and identified the maximizing weights of the empirical likelihood that carry the correlation information induced via the matrix expansion ideas of the quadratic inference function. In the second step they incorporated the weights in the quantile marginal regression to increase efficiency. This indirect two-step approach, however, ignores the fact that the correlation structure involved in the quantile regression is the sign correlation, whereas it is the standard Pearson correlation that is involved with the conditional mean regression. Therefore, the correlation information incorporated as weights stays the same regardless of the conditional regression quantile of interest. The proposed approach induces the correlation information specific to the regression quantile of interest directly. The matrix expansion ideas of the quadratic inference function alone was considered by Leng and Zhang (2014) without accounting for dropouts.

As far as the treatment of dropouts is concerned, the proposed procedure is similar to those of Lipsitz *and others* (1997), Robins *and others* (1995), and Yi and He (2009). The proposed procedure differs in that it accommodates the correlated nature of the longitudinal data using an informative working correlation structure without requiring the exact knowledge of the true correlation structure nor estimating the informative working correlation structure. The existing methods either ignore the within-subject correlation structure (Robins *and others*, 1995; Lipsitz *and others*, 1997) or require the correlation structure to be estimated (Yi and He, 2009).

The remainder of this paper proceeds as follows: in Section 2, we first propose the new marginal quantile regression procedure for longitudinal data, and then expand it to account for dropouts that are missing at random. In Section 3, we illustrate the methodology using both simulation studies and a real-life data analysis of an HIV study. We conclude the paper with some final remarks in Section 4.

## 2. Methodology

### 2.1 *Quantile marginal regression with an informative working correlation structure*

Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})^{\mathrm{T}}$ be the $i$th subject's responses for $i = 1, \ldots, n$, where $n$ is the sample size and $m$ is the number of longitudinal measurements taken on each subject over time. Given $\tau \in (0, 1)$, the $\tau$th quantile marginal regression model for the longitudinal data is formulated as $\mathbf{y}_i = \mathbf{x}_i \boldsymbol{\beta}_\tau + \boldsymbol{\epsilon}_i$, where $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im})^{\mathrm{T}}$ is an $(m \times p)$-dimensional matrix of covariate, $\boldsymbol{\beta}_\tau$ is a true parameter vector, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{im})^{\mathrm{T}}$ is the random error vector satisfying $P(\epsilon_{ij} < 0 \,|\, \mathbf{x}_{ij}) = \tau$ for any $i$ and $j$. If repeated measurements of each subject are assumed independent, an estimator of $\boldsymbol{\beta}_\tau$ is obtained by minimizing the following objective function:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \rho_\tau (y_{ij} - \mathbf{x}_{ij} \boldsymbol{\beta}), \tag{2.1}$$

where $\rho_\tau(u) = u\{\tau - 1(u < 0)\}$ is the so-called check function and $1(\cdot)$ is the indicator function. We let $\varphi_\tau(u) = \rho'_\tau(u)$ for $u \neq 0$ and $\varphi_\tau(u) = 0$ otherwise. From (2.1), estimating equations can be derived by differentiating $S(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ as follows:

$$\sum_{i=1}^{n} \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\varphi}_\tau (\mathbf{y}_i - \mathbf{x}_i \boldsymbol{\beta}) = 0. \tag{2.2}$$

Due to the discontinuity of $\boldsymbol{\varphi}_\tau(\mathbf{y}_i - \mathbf{x}_i \boldsymbol{\beta}) = \{\varphi_\tau(y_{i1} - \mathbf{x}_{i1} \boldsymbol{\beta}), \ldots, \varphi_\tau(y_{im} - \mathbf{x}_{im} \boldsymbol{\beta})\}^{\mathrm{T}}$, an estimator of $\boldsymbol{\beta}_\tau$ may only satisfy the equations approximately. This modeling ignores the correlated nature of longitudinal data and may cause a loss of efficiency.

In order to incorporate the within-subject correlation information, we extend the estimating equations as

$$\sum_{i=1}^{n} \mathbf{x}_i^{\mathrm{T}} \mathbf{V}_i^{-1} \boldsymbol{\varphi}_\tau (\mathbf{y}_i - \mathbf{x}_i \boldsymbol{\beta}) = 0, \tag{2.3}$$

where $\mathbf{V}_i$ is the covariance matrix of $\boldsymbol{\varphi}_\tau(\mathbf{y}_i - \mathbf{x}_i \boldsymbol{\beta}_\tau)$. The inverse of the covariance matrix $\mathbf{V}_i^{-1}$ can be decomposed as $\mathbf{A}_i^{-1/2} \boldsymbol{\Phi}_i^{-1} \mathbf{A}_i^{-1/2}$, with $\mathbf{A}_i = \mathrm{diag}(a_{i1}, \ldots, a_{im})$ being a $(m \times m)$-dimensional diagonal marginal variance matrix of $\boldsymbol{\varphi}_\tau(\mathbf{y}_i - \mathbf{x}_i \boldsymbol{\beta}_\tau)$ and $\boldsymbol{\Phi}_i$ being an $(m \times m)$-dimensional true correlation matrix. In practice, $\boldsymbol{\Phi}_i$ is unknown and we utilize a working correlation structure (denoted by $\mathbf{R}_i$). The $j$th element

of $\mathbf{A}_i$ is $a_{ij} = \text{var}\{\varphi_\tau(y_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta}_\tau)\} = \tau(1-\tau)$ for all $j$. Thus, given $\mathbf{R}_i$, equation (2.3) can be simplified as

$$\sum_{i=1}^{n} \mathbf{x}_i^{\mathrm{T}} \mathbf{R}_i^{-1} \boldsymbol{\varphi}_\tau(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}) = 0. \tag{2.4}$$

Motivated by Qu *and others* (2000), we represent the inverse of the working correlation $\mathbf{R}_i^{-1}$ in (2.4) by a linear combination of basis matrices, $\mathbf{R}_i^{-1} = \sum_{j=1}^{q} b_j \mathbf{B}_{ij}$, where $\mathbf{B}_{i1}, \ldots, \mathbf{B}_{iq}$ are $(m \times m)$-dimensional basis matrices depending on the particular choice of $\mathbf{R}$ and $b_1, \ldots, b_q$ are unknown coefficients. For example, if a working correlation structure is the compound symmetry, then $\mathbf{R}_i^{-1} = b_1\mathbf{B}_{i1} + b_2\mathbf{B}_{i2}$, where $\mathbf{B}_{i1}$ is an identity matrix and $\mathbf{B}_{i2}$ is a symmetric matrix with 0 on the diagonal and 1 elsewhere. The coefficients $b_0$ and $b_1$ are associated with the compound symmetry correlation parameter. If $\mathbf{R}_i$ corresponds to AR(1), $\mathbf{R}_i^{-1} = b_1\mathbf{B}_{i1} + b_2\mathbf{B}_{i2} + b_3\mathbf{B}_{i3}$, where $\mathbf{B}_{i1}$ is an identity matrix, $\mathbf{B}_{i2}$ is a symmetric matrix with 1 on the sub-diagonal entries and 0 elsewhere, and $\mathbf{B}_{i3}$ is a symmetric matrix with 1 in elements (1, 1) and $(m, m)$, and 0 elsewhere with corresponding coefficients $b_1$, $b_2$, and $b_3$, respectively. In general, $\mathbf{B}_{i3}$ is a minor boundary correction and can be omitted. More details are also provided in Qu *and others* (2000) and Cho and Qu (2015).

Consequently, equation (2.4) can be approximated as a linear combination of the elements, $\mathbf{g}_i(\boldsymbol{\beta})$ for $i = 1, \ldots, n$, where

$$\mathbf{g}_i(\boldsymbol{\beta}) = \begin{pmatrix} \mathbf{x}_i^{\mathrm{T}}\mathbf{B}_{i1}\boldsymbol{\varphi}_\tau(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}) \\ \vdots \\ \mathbf{x}_i^{\mathrm{T}}\mathbf{B}_{iq}\boldsymbol{\varphi}_\tau(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}) \end{pmatrix}. \tag{2.5}$$

If $E\{\mathbf{g}_i(\boldsymbol{\beta}_\tau)\} = 0$, the following empirical likelihood function can be constructed for the inference of $\boldsymbol{\beta}_\tau$:

$$L(\boldsymbol{\beta}) = \max\left\{ \prod_{i=1}^{n} p_i \,\middle|\, \sum_{i=1}^{n} p_i \mathbf{g}_i(\boldsymbol{\beta}) = 0, \sum_{i=1}^{n} p_i = 1, 0 \leqslant p_i \leqslant 1 \right\}, \tag{2.6}$$

where $p_i$ denotes a point mass assigned to the $i$th data point $(\mathbf{x}_i, \mathbf{y}_i)$. We consider a set of weights $\{\hat{p}_i\}_{i=1}^n$ that maximizes the empirical likelihood in equation (2.6). Following Qin and Lawless (1994), we define the maximum empirical likelihood estimator $\hat{\boldsymbol{\beta}}_\tau$ as a solution to the equation $\sum_{i=1}^n \hat{p}_i \mathbf{g}_i(\boldsymbol{\beta}) = 0$, or alternatively as the maximizer of $L(\boldsymbol{\beta})$:

$$\hat{\boldsymbol{\beta}}_\tau = \underset{\boldsymbol{\beta}}{\text{argmax}}\, L(\boldsymbol{\beta}). \tag{2.7}$$

Note that estimation of the parameters $b_1, \ldots, b_q$ is not required, since the function $\mathbf{g}_i(\boldsymbol{\beta})$ does not involve the parameters. Also note that $\mathbf{g}_i(\boldsymbol{\beta})$ is a $p \times q$ variate function, and hence the constraint $\sum_{i=1}^n p_i \mathbf{g}_i(\boldsymbol{\beta}) = 0$ is an overdetermined system. The maximization in (2.7) can be conducted by the existing R package *optim* with the empirical likelihood (2.6) as the objective function. Given $\boldsymbol{\beta}$, the empirical likelihood is evaluated by the R package *emplik*. We defer discussion of computational details to Section 3.2.

We assume the following conditions to study the asymptotic properties of $\hat{\boldsymbol{\beta}}_\tau$:

CONDITION 1 $E\{\mathbf{g}_i(\boldsymbol{\beta}_\tau)\} = 0$ and $E[\mathbf{g}_i(\boldsymbol{\beta}_\tau)\{\mathbf{g}_i(\boldsymbol{\beta}_\tau)\}^{\mathrm{T}}]$ are positive definite.

CONDITION 2 Let $F_{ij}(\cdot \mid x_{ij})$ denote the cumulative distribution function of $e_{ij}$ given $\mathbf{x}_{ij}$. We see that $F_{ij}$ is twice continuously differentiable with derivatives bounded in the neighborhood of zeros uniformly in $\mathbf{x}_{ij}$ for all $i$ and $j$.

CONDITION 3 The random vectors $\mathbf{x}_{ij}$ are bounded in probability for all $i$ and $j$.

These are standard conditions commonly assumed for quantile regression. We define $\mathbf{B}_{ij}^{\mathrm{T}}\mathbf{x}_i = \mathbf{Z}_{i(j)}$ and $f_{ij}(\cdot \mid x_{ij}) = F'_{ij}(\cdot \mid x_{ij})$. Given $\tau$, we simplify the notations by letting $\boldsymbol{\varphi}_i(\boldsymbol{\beta}) = \boldsymbol{\varphi}_\tau(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta})$. Then, $\mathbf{g}_i^{\mathrm{T}}(\boldsymbol{\beta}) = (\boldsymbol{\varphi}_i^{\mathrm{T}}(\boldsymbol{\beta})\mathbf{Z}_{i(1)}, \ldots, \boldsymbol{\varphi}_i^{\mathrm{T}}(\boldsymbol{\beta})\mathbf{Z}_{i(q)})$. We further define

$$\boldsymbol{\Delta}_i = \mathrm{diag}\{f_{i1}(0 \mid \mathbf{x}_{i1}), \ldots, f_{in_i}(0 \mid \mathbf{x}_{in_i})\}, \quad \tilde{\mathbf{D}}_1 = E(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\Delta}_i\mathbf{x}_i), \quad \tilde{\mathbf{D}}_0 = E(\mathbf{x}_i^{\mathrm{T}}\Phi_i\mathbf{x}_i),$$
$$\mathbf{D}_1^{\mathrm{T}} = E(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\Delta}_i\mathbf{Z}_{i(1)}, \ldots \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\Delta}_i\mathbf{Z}_{i(q)}), \quad \mathbf{D}_0 = E\{\mathbf{g}_i(\boldsymbol{\beta}_\tau)\mathbf{g}_i^{\mathrm{T}}(\boldsymbol{\beta}_\tau)\}.$$

We note that $\tilde{\mathbf{D}}_0$, $\tilde{\mathbf{D}}_1$, and $\mathbf{D}_0$ are positive definite and the rank of $\mathbf{D}_1$ is $p$ under conditions 1–3. We adopt the notation $\mathbf{M}_1 \leqslant \mathbf{M}_2$ for square matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ of the same order, when $\mathbf{M}_2 - \mathbf{M}_1$ is positive semidefinite.

THEOREM 2.1 Assume conditions 1–3 hold. Given $\mathbf{R}_i$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau) \longrightarrow N(0, \tau(1-\tau)\mathbf{V})$$

as $n \to \infty$, where $\mathbf{V} = \{\mathbf{D}_1^{\mathrm{T}}\mathbf{D}_0^{-1}\mathbf{D}_1\}^{-1}$. Furthermore, $\mathbf{V} \leqslant \tilde{\mathbf{V}}$ for $\tilde{\mathbf{V}} = \{\tilde{\mathbf{D}}_1^{\mathrm{T}}\tilde{\mathbf{D}}_0^{-1}\tilde{\mathbf{D}}_1\}^{-1}$ with the equality holding if $\mathbf{R}_i$ is an $(n_i \times n_i)$-dimensional identity matrix $\mathbf{I}$.

As shown in He *and others* (2003), $\tau(1-\tau)\tilde{\mathbf{V}}$ is the asymptotic variance if $\mathbf{R}_i = \mathbf{I}$, i.e. under the working independence assumption. As $\mathbf{V} \leqslant \tilde{\mathbf{V}}$, the asymptotic variance under $\mathbf{R}_i \neq \mathbf{I}$ is no greater than the asymptotic variance obtained under working independence. We see that $\tilde{\mathbf{V}} - \mathbf{V}$ accounts for the efficiency gain from incorporating the within-subject dependency commonly existing in the longitudinal data. Importantly, the efficiency gain does not require that the assumed working correlation structure be correctly specified.

### 2.2 *Weighted quantile marginal regression with dropouts*

In longitudinal studies, subjects may drop out of the study before the end of the follow-up, and the number of observed within-subject measurements (denoted by $m_i$) may vary. If the missingness is not associated with the data, we may ignore missing responses and apply the empirical likelihood procedure developed in Section 2.1 to analyze the observed data. However, if the missingness is related to the covariates or observed responses, ignoring missing responses often results in bias. In this section, we extend the empirical likelihood procedure to accommodate missing responses that depends on the observed data.

We denote $v_{ij}$ as the observed data for the $i$th subject at time $j$, which potentially could include covariates $\mathbf{x}_i$ and observed responses $y_{ik}$ up to time $j$ $(k < j)$. Let $M_{ij}$ be a missing indicator variable being 0 if missing and 1 otherwise, and assume that all individuals are observed at the first assessment, i.e. $M_{i1} = 1$ for all $i = 1, \ldots, n$. We assume that the responses are missing at random conditioning on the observed data $v_{ij}$, so the probability of not dropping out at time $j$ is given by $\pi_{ij} = P(M_{ij} = 1 \mid v_{ij})$. We propose the following weighted estimating equation:

$$\sum_{i=1}^{n} \mathbf{x}_i^{\mathrm{T}}\mathbf{R}_i^{-1}\mathbf{W}_i\boldsymbol{\varphi}_\tau(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_\tau) = 0, \tag{2.8}$$

where $\mathbf{W}_i = \text{diag}(M_{i1}/\pi_{i1}, \ldots, M_{im_i}/\pi_{im_i})$. Based on the weighted estimating equation (2.8), the estimating function in (2.5) is accordingly modified as

$$\mathbf{g}_i^w(\boldsymbol{\beta}) = \begin{pmatrix} \mathbf{x}_i^{\mathrm{T}}\mathbf{B}_{i1}\mathbf{W}_i\boldsymbol{\varphi}_\tau(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}) \\ \vdots \\ \mathbf{x}_i^{\mathrm{T}}\mathbf{B}_{iq}\mathbf{W}_i\boldsymbol{\varphi}_\tau(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}) \end{pmatrix}. \tag{2.9}$$

Since $\pi_{ij}$ is often unknown in practice, $\pi_{ij}$ should be estimated and substituted in (2.9). We consider consistent estimators of $\pi_{ij}$ and obtain $\hat{\mathbf{g}}_i^w(\boldsymbol{\beta})$ by replacing $\mathbf{W}_i$ in (2.9) with $\widehat{\mathbf{W}}_i = \text{diag}(M_{i1}/\hat{\pi}_{i1}, \ldots, M_{im_i}/\hat{\pi}_{im_i})$. We have the following empirical likelihood for the inference of $\boldsymbol{\beta}_\tau$ under some regularity conditions:

$$L^w(\boldsymbol{\beta}) = \max\left\{ \prod_{i=1}^n p_i \,\middle|\, \sum_{i=1}^n p_i \widehat{\mathbf{g}}_i^w(\boldsymbol{\beta}) = 0, \sum_{i=1}^n p_i = 1, 0 \leqslant p_i \leqslant 1 \right\}. \tag{2.10}$$

Similarly, following Qin and Lawless (1994), we consider a set of weights $\{\hat{p}_i^w\}_{i=1}^n$ that maximizes (2.10) and define the maximum empirical likelihood estimator $\hat{\boldsymbol{\beta}}_\tau^w$ as a solution to the equation, $\sum_{i=1}^n \hat{p}_i^w \hat{\mathbf{g}}_i^w(\boldsymbol{\beta}) = 0$, or alternatively as the maximizer

$$\hat{\boldsymbol{\beta}}_\tau^w = \underset{\boldsymbol{\beta}}{\text{argmax}}\, L^w(\boldsymbol{\beta}). \tag{2.11}$$

For the estimation of $\pi_{ij}$, we assume a parametric model and denote the unknown vector of the parameter model by $\boldsymbol{\alpha}$. We let $\boldsymbol{\alpha}_0$ denote the true value of $\boldsymbol{\alpha}$ and $S_i(\boldsymbol{\alpha})$ the score function. An estimator is obtained from solving $\sum_{i=1}^n S_i(\boldsymbol{\alpha}) = 0$ and we denote it by $\hat{\boldsymbol{\alpha}}$. Details of modeling and computation are provided in Sections 3.1 and 3.2 when logistic regression is assumed. We denote $\partial S_i(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$ evaluated at $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ by $\partial S_i(\boldsymbol{\alpha}_0)/\partial\boldsymbol{\alpha}$ and adopt this convention throughout the paper. As $\mathbf{W}_i = \mathbf{W}_i(\boldsymbol{\alpha})$, we have $\mathbf{g}_i^w(\boldsymbol{\beta}) = \mathbf{g}_i^w(\boldsymbol{\alpha}, \boldsymbol{\beta})$. We define

$$\mathbf{D}_1^{w\mathrm{T}} = E\{\partial \mathbf{g}_i^w(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_\tau)/\partial\boldsymbol{\beta}\}, \quad \mathbf{D}_0^w = E[\mathbf{g}_i^w(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_\tau)\{\mathbf{g}_i^w(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_\tau)\}^{\mathrm{T}}],$$
$$\mathbf{D}_s = E[\{\partial \mathbf{g}_i^w(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_\tau)\}/\partial\boldsymbol{\alpha}], \quad \mathbf{U}_i = \mathbf{g}_i^w(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_\tau) - \mathbf{D}_s E\{\partial S_i(\boldsymbol{\alpha}_0)/\partial\boldsymbol{\alpha}\}^{-1} S_i(\boldsymbol{\alpha}_0).$$

THEOREM 2.2 Assume conditions 1–3. Suppose that $S_i(\boldsymbol{\alpha})$ is a continuous function and $\pi_{ij}$ are bounded below from zero uniformly in $\nu_{ij}$ for all $i$ and $j$. We also suppose the random vectors $\nu_{ij}$ are bounded in probability for all $i$ and $j$. Given a working correlation matrix $\mathbf{R}_i$, as $n \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_\tau^w - \boldsymbol{\beta}_\tau) \longrightarrow N(0, \tau(1-\tau)\mathbf{V}^w),$$

where $\mathbf{V}^w = \{\mathbf{D}_1^{w\mathrm{T}}(\mathbf{D}_0^w)^{-1}\mathbf{D}_1^w\}^{-1}\mathbf{D}_1^{w\mathrm{T}}(\mathbf{D}_0^w)^{-1}\{E(\mathbf{U}_i\mathbf{U}_i^{\mathrm{T}})\}(\mathbf{D}_0^w)^{-1}\mathbf{D}_1^w\{\mathbf{D}_1^{w\mathrm{T}}(\mathbf{D}_0^w)^{-1}\mathbf{D}_1^w\}^{-1}$.

If $\pi_{ij}$ were known, $\mathbf{V}^w$ would be reduced to $\{\mathbf{D}_1^{w\mathrm{T}}(\mathbf{D}_0^w)^{-1}\mathbf{D}_1^w\}^{-1}$, which is similar to the corresponding quantity $\mathbf{V}$ in the complete case with no missingness. Therefore, we note that unknown $\pi_{ij}$ would add to the complexity in the form of $\mathbf{V}^w$.

With $\mathbf{R}_i = \mathbf{I}$ the asymptotic variance becomes $\tau(1-\tau)\tilde{\mathbf{V}}^w$, where

$$\tilde{\mathbf{V}}^w = \{\tilde{\mathbf{D}}_1^{w\,\mathrm{T}}(\tilde{\mathbf{D}}_0^w)^{-1}\tilde{\mathbf{D}}_1^w\}^{-1}\tilde{\mathbf{D}}_1^{w\,\mathrm{T}}(\tilde{\mathbf{D}}_0^w)^{-1}\{E(\tilde{\mathbf{U}}_i\tilde{\mathbf{U}}_i^{\mathrm{T}})\}(\tilde{\mathbf{D}}_0^w)^{-1}\tilde{\mathbf{D}}_1^w\{\tilde{\mathbf{D}}_1^{w\,\mathrm{T}}(\tilde{\mathbf{D}}_0^w)^{-1}\tilde{\mathbf{D}}_1^w\}^{-1} \quad \text{with}$$

$$\tilde{\mathbf{D}}_1^w = E\{\mathbf{x}_i^{\mathrm{T}}\mathbf{W}_i(\boldsymbol{\alpha}_0)\boldsymbol{\Delta}_i\mathbf{x}_i\}, \quad \tilde{\mathbf{D}}_0^w = E\{\mathbf{x}_i^{\mathrm{T}}\mathbf{w}_i(\boldsymbol{\alpha}_0)\boldsymbol{\Phi}_i\mathbf{W}_i(\boldsymbol{\alpha}_0)\mathbf{x}_i\},$$

$$\tilde{\mathbf{D}}_s = E[\{\partial\mathbf{x}_i^{\mathrm{T}}\mathbf{W}_i(\boldsymbol{\alpha}_0)\boldsymbol{\varphi}_i(\boldsymbol{\beta}_\tau)\}/\partial\boldsymbol{\alpha}], \quad \tilde{\mathbf{U}}_i = \mathbf{x}_i^{\mathrm{T}}\mathbf{W}_i(\boldsymbol{\alpha}_0)\boldsymbol{\varphi}_i(\boldsymbol{\beta}_\tau) - \tilde{\mathbf{D}}_s E\{\partial S_i(\boldsymbol{\alpha}_0)/\partial\boldsymbol{\alpha}\}^{-1}S_i(\boldsymbol{\alpha}_0).$$

For known $\pi_{ij}$, $\tilde{\mathbf{V}}^w = \left\{\tilde{\mathbf{D}}_1^{w\mathrm{T}}(\tilde{\mathbf{D}}_0^w)^{-1}\tilde{\mathbf{D}}_1^w\right\}^{-1}$ and it can be easily shown that $\mathbf{V}^w \leqslant \tilde{\mathbf{V}}^w$. This suggests that incorporating the within-subject correlation results in efficiency gain as in the no-dropout case. With the estimated $\pi_{ij}$, it is not straightforward to obtain analytic results; however, empirical studies in Section 3 show that choosing proper working correlation structures is still beneficial.

## 3. Empirical studies

In this section, we evaluate performance of the proposed method using simulation studies and a real-life data analysis example. The simulation setups reflect the real-life data example in order to provide estimates of the operating characteristics of the proposed methods in the real-life example. We include various cases of error distributions; however, and the simulation results are generalizable.

### 3.1 *Simulation studies*

We generate data from the following regression model:

$$y_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2j} + \beta_3 x_{i1}x_{i2j} + \epsilon_{ij}, \text{ for } \quad i = 1, \ldots, 200 \quad \text{and} \quad j = 1, \ldots, m_i,$$

where $(\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}} = (6, -1, -1, 0.5)^{\mathrm{T}}$, $x_{i1}$ are treatment indicators generated from a Bernoulli distribution with $P(x_{i1} = 1) = 0.5$ that reflects $1:1$ randomization, $x_{i2j}$ indicate the follow-up measurement times $j$, and $x_{i1}x_{i2j}$ are the interaction terms between treatment and time effect. We explore the following three distributions for the random error $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{im_i})^{\mathrm{T}}$:

*Case* 1. Normal errors: $\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is an AR(1) correlation structure with correlation coefficients of 0.7.
*Case* 2. Asymmetric errors: $\boldsymbol{\epsilon}_i = exp(\boldsymbol{\zeta}_i) - 1$, where $\boldsymbol{\zeta}_i \sim N(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is defined in Case 1.
*Case* 3. Heteroskedastic errors: $\boldsymbol{\epsilon}_i = (1 + \mathbf{x}_{i2}/10)\boldsymbol{\zeta}_i$, where $\boldsymbol{\zeta}_i$ is specified in Case 2.

We assess the regression quantile at $\tau = 0.25, 0.5$, and $0.75$ from 500 simulated datasets, and explore three common working correlation choices: independence, AR(1), and compound symmetry. The within-subject correlations involved with the quantile regression are sign correlations, i.e. $\mathrm{cor}\{\tau - 1(\epsilon_{ij} < 0), \tau - 1(\epsilon_{ik} < 0)\}$ for $j, k = 1, \ldots, m_i$. Hence the true correlation structure is a toeplitz with $(m_i - 1)$ number of parameters $\rho_{|j-k|} = \mathrm{cor}\{\tau - 1(\epsilon_{ij} < 0), \tau - 1(\epsilon_{ik} < 0)\}$ for $j \neq k$. Among the three common working correlation choices (i.e. AR(1), compound symmetry, exchangeable), the AR(1) structure best approximates the true correlation structure.

To generate longitudinal data with dropouts, we assumed that each subject is repeatedly measured five times with equally spaced time, and then indicated dropouts as $M_{ij}$ with $M_{ij} = 0$ if missing and $M_{ij'} = 0$ for all $j' > j$. Let $\lambda_{ij} = P(M_{ij} = 1 \mid M_{i,j-1} = 1, v_{ij})$ with $v_{ij} = (x_{i2j}, y_{i,j-1})$. To model the dropout process, we adapt the commonly used logistic regression models as follows:

$$\mathrm{logit}\,\lambda_{ij} = \alpha_1 x_{i2j} + \alpha_2 y_{i,j-1}, \tag{3.1}$$

Table 1. *Mean squared error*, *bias*, *and coverage probabilities of the proposed approach in case of data with dropouts under different assumptions on the dropouts* (*missing at random* (*MAR*) *and missing completely at random* (*MCAR*)) *with three working correlation structures* (*AR*(1), *compound symmetry* (*CS*) *and independence* (*IND*)) *at* $\tau = 0.5$

| | | Bias | | | | Coverage | | | |
|---|---|---|---|---|---|---|---|---|---|
| | MSE | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_0$ (%) | $\beta_1$ (%) | $\beta_2$ (%) | $\beta_3$(%) |
| Case 1 | | | | | | | | | |
| MAR | | | | | | | | | |
| AR(1) | 0.012 | −0.002 | 0.016 | 0.004 | −0.013 | 94 | 93 | 93 | 94 |
| CS | 0.015 | 0.001 | 0.024 | 0.004 | −0.017 | 94 | 92 | 94 | 93 |
| IND | 0.017 | −0.001 | 0.027 | 0.007 | −0.018 | 92 | 91 | 93 | 93 |
| MCAR | | | | | | | | | |
| AR(1) | 0.017 | −0.070 | 0.021 | 0.069 | −0.014 | 84 | 90 | 56 | 94 |
| CS | 0.019 | −0.097 | 0.027 | 0.080 | −0.021 | 85 | 90 | 58 | 94 |
| IND | 0.021 | −0.101 | 0.019 | 0.081 | −0.018 | 83 | 92 | 53 | 94 |
| Case 2 | | | | | | | | | |
| MAR | | | | | | | | | |
| AR(1) | 0.013 | 0.009 | 0.001 | −0.003 | −0.002 | 94 | 93 | 93 | 97 |
| CS | 0.016 | 0.011 | 0.005 | 0.002 | −0.004 | 94 | 93 | 98 | 96 |
| IND | 0.017 | 0.024 | 0.002 | 0.002 | −0.002 | 92 | 91 | 94 | 93 |
| MCAR | | | | | | | | | |
| AR(1) | 0.017 | −0.025 | −0.058 | 0.057 | 0.016 | 94 | 89 | 84 | 94 |
| CS | 0.020 | −0.067 | −0.047 | 0.071 | 0.018 | 92 | 94 | 79 | 96 |
| IND | 0.022 | −0.086 | −0.046 | 0.078 | 0.021 | 94 | 96 | 78 | 96 |
| Case 3 | | | | | | | | | |
| MAR | | | | | | | | | |
| AR(1) | 0.018 | −0.012 | 0.021 | 0.018 | −0.015 | 94 | 94 | 93 | 95 |
| CS | 0.029 | −0.024 | 0.031 | 0.020 | −0.018 | 92 | 92 | 91 | 94 |
| IND | 0.032 | −0.030 | 0.022 | 0.022 | −0.016 | 91 | 91 | 90 | 92 |
| MCAR | | | | | | | | | |
| AR(1) | 0.028 | −0.116 | 0.019 | 0.118 | −0.010 | 83 | 88 | 49 | 90 |
| CS | 0.036 | −0.152 | 0.019 | 0.136 | −0.012 | 81 | 91 | 46 | 92 |
| IND | 0.038 | −0.181 | 0.012 | 0.145 | −0.009 | 78 | 91 | 35 | 92 |

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^{\mathrm{T}} = (-0.6, 1)^{\mathrm{T}}$. The negative value of $\alpha_1$ indicates that a subject is more likely to drop out from the study over time. The positive coefficient of $\alpha_2$ implies that the larger the previously measured response values are, the less likely subjects are to dropout. Note that the probability $\pi_{ij} = P(M_{ij} = 1 \mid \nu_{ij})$ can be expressed as $\pi_{ij} = \prod_{t=2}^{j} \lambda_{it}$. Consequently, the estimate of $\lambda_{ij}$ is obtained as $\hat{\lambda}_{ij} = 1/(1 + e^{-\hat{\alpha}_1 x_{i2j} - \hat{\alpha}_2 y_{i,j-1}})$. In this setup, the percentage of patients dropping out increases gradually from 4.3% on average after the first visit and the total percentage of dropouts at the last visit reaches 74.0% on average when the normal error distribution was considered (Case 1). Dropouts patterns and the percentages of dropouts are similar with the asymmetric errors (Case 2) and heteroskedastic errors (Case 3).

With $\hat{\mathbf{W}}_i = \mathrm{diag}(M_{i1}/\hat{\pi}_{i1}, \ldots, M_{im_i}/\hat{\pi}_{im_i})$ given by $\hat{\pi}_{i1} = 1$ and $\hat{\pi}_{ij} = \prod_{t=2}^{j} \hat{\lambda}_{it}$ for $j > 1$, we implemented the weighted empirical likelihood approach in (2.10). We also considered the unweighted approach. This corresponds to the case in which missingness is assumed completely at random, whereas the missing mechanism described in (3.1) is missing at random. Table 1 presents the mean square error estimates of

Table 2. *Mean squared error estimates in case of no dropouts using methods by* Tang and Leng (2011) *and* Yi and He (2009), *and the proposed approach. Working correlation structures denoted by AR*(1) *and CS correspond to autoregressive correlation structure of order* 1 *and compound symmetry and independence structure*, *respectively*

| | | Proposed method | | Tang and Leng | | |
|---|---|---|---|---|---|---|
| Case | Quantile | AR(1) | CS | AR(1) | CS | Yi and He |
| 1 | 0.25 | 0.010 | 0.012 | 0.020 | 0.021 | 0.013 |
| | 0.5 | 0.009 | 0.010 | 0.018 | 0.018 | 0.013 |
| | 0.75 | 0.010 | 0.013 | 0.020 | 0.020 | 0.017 |
| 2 | 0.25 | 0.004 | 0.004 | 0.006 | 0.006 | 0.004 |
| | 0.5 | 0.012 | 0.014 | 0.018 | 0.020 | 0.016 |
| | 0.75 | 0.044 | 0.058 | 0.094 | 0.096 | 0.088 |
| 3 | 0.25 | 0.013 | 0.014 | 0.028 | 0.028 | 0.018 |
| | 0.5 | 0.012 | 0.014 | 0.022 | 0.023 | 0.018 |
| | 0.75 | 0.015 | 0.019 | 0.026 | 0.026 | 0.024 |

$\hat{\boldsymbol{\beta}}^{w}_{\tau=0.5}$ and the average of bias estimates of each coefficient. The coverage probability of bootstrap confidence intervals at the nominal 95% level was also reported. The confidence intervals were obtained from 200 reputations.

Among the results obtained under the missing at random assumption, the mean square errors were smallest under the AR(1) working correlation structure and the largest under the independent working correlation structure for all cases. This confirms that we obtained a more efficient estimator by assuming an informative working correlation structure when compared with ignoring the correlation structure, even though the specified correlation structure may not be correct. Moreover, bias was smaller and the coverage probabilities were closer to the nominal 95% level with the informative working correlation structures. On the other hand, when missingness was erroneously assumed completely at random, the estimators were no longer unbiased and most coverage probabilities were below the nominal level. The coverage probabilities of $\beta_2$ in Case 3 were even smaller than 50%. Results at $\tau = 0.25$ and 0.75 were similar and provided in supplementary materials (available at *Biostatistics* online).

Furthermore, we assumed no dropouts ($m_i = 5$ for all subjects) and compared the proposed method in (2.6) with those of Yi and He (2009) and Tang and Leng (2011). Although all yielding asymptotically unbiased estimators, Table 2 reports that the proposed empirical likelihood method yielded a more efficient estimator in all the finite sample cases under consideration.

### 3.2 *Real-life data application*

The motivating dataset comes from ACTG116 study by Dolin *and others* (1995). This is a longitudinal controlled trial of HIV disease in patients with advanced HIV Type 1 infection. It measured CD4 cell counts repeatedly in each patient at weeks 0, 16, 32, 48, and 64, and aimed to assess treatment effects on CD4 cell counts with possible time effects. CD4 cell counts are a biomarker for AIDS or AIDS-related complex diseases. They generally decrease as the HIV patient's immune system deteriorates. We analyzed CD4 cell counts data on 408 patients who were randomly assigned to the following two treatment groups: zidovudine (211 subjects) and didanosine (197 subjects).

Figure 1 indicates that the distribution of CD4 cell counts are skewed. The mean regression may not appropriately assess the longitudinal change in the CD4 cell count. As an alternative, we postulated the

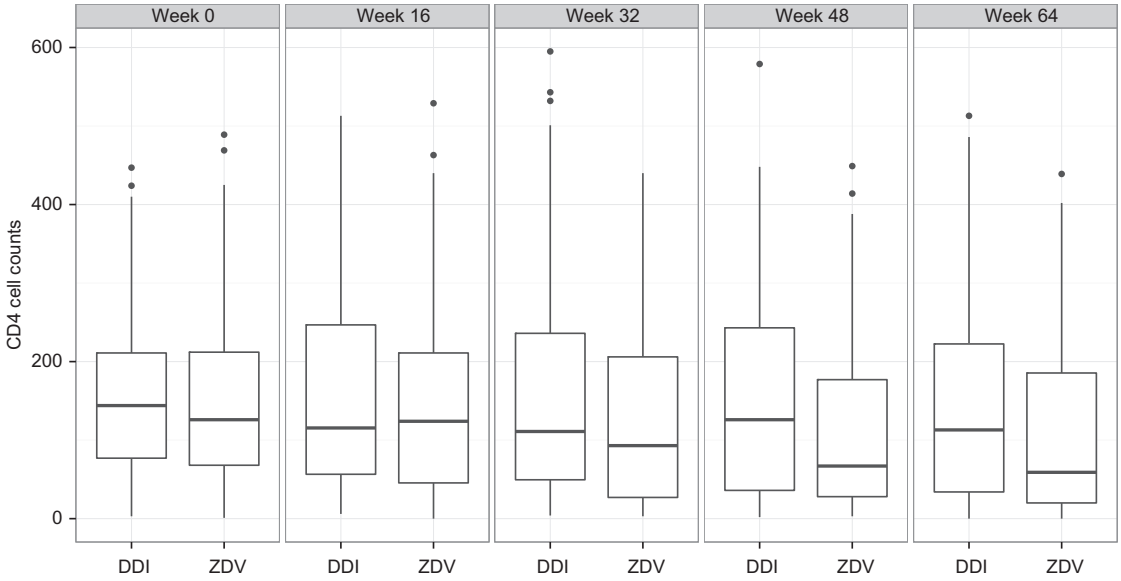


Fig. 1. Boxplots of CD4 cell counts for didanosine (DDI) and zidovudine (ZDV) groups at weeks 0, 16, 32, 48, and 64.

Table 3. *Estimated coefficients with the standard errors*: *a coefficient whose bootstrap confidence interval does not include zero is marked with* $*$ *and the generalized estimating equation approach is denoted by GEE*

| Conditional regression | Method | $\text{Intercept}_{se}$ | $\text{Treatment}_{se}$ | $\text{Time}_{se}$ | $\text{Interaction}_{se}$ |
|---|---|---|---|---|---|
| Mean | GEE-AR(1) | $154.03_{7.17}*$ | $1.77_{10.26}$ | $-5.87_{2.57}*$ | $-6.56_{3.29}*$ |
| | GEE-independence | $151.11_{7.95}*$ | $5.13_{11.12}$ | $0.26_{3.23}$ | $-10.32_{4.21}*$ |
| Median | Proposed | $140.63_{5.85}*$ | $9.29_{6.90}$ | $-15.54_{2.54}*$ | $-5.83_{3.24}$ |
| | Naive 1 | $136.07_{6.02}*$ | $12.39_{7.78}$ | $-8.86_{5.15}$ | $-11.83_{4.81}*$ |
| | Naive 2 | $137.92_{8.97}*$ | $11.21_{12.14}$ | $-5.96_{5.09}$ | $-12.77_{5.86}*$ |
| 0.25th quantile | Proposed | $76.18_{2.61}*$ | $-3.54_{3.72}$ | $-10.91_{0.95}*$ | $-1.04_{1.07}$ |
| | Naive 1 | $78.04_{4.50}*$ | $0.12_{4.66}$ | $-8.58_{1.22}*$ | $-2.02_{1.48}$ |
| | Naive 2 | $77.00_{5.87}*$ | $-2.00_{6.69}$ | $-9.00_{1.59}*$ | $-3.00_{1.83}$ |
| 0.75th quantile | Proposed | $217.37_{7.25}*$ | $29.26_{8.45}*$ | $-2.43_{3.60}$ | $-15.38_{5.48}*$ |
| | Naive 1 | $207.64_{7.66}*$ | $21.41_{9.37}*$ | $5.62_{4.96}$ | $-19.83_{6.82}*$ |
| | Naive 2 | $215.02_{10.72}*$ | $12.24_{13.81}$ | $5.09_{4.46}$ | $-13.75_{7.11}$ |

following quantile regression model:

$$Q_\tau(x_{ij}) = \beta_{0,\tau} + \beta_{1,\tau}x_{i1j} + \beta_{2,\tau}x_{i2j} + \beta_{3,\tau}x_{i1}x_{i2j}, \quad \text{for } i = 1, \ldots, 408 \quad \text{and} \quad j = 1, \ldots, m_i,$$

where $Q_\tau(x_{ij})$ denotes the $\tau$th conditional quantile of the CD4 cell count given the covariates $x_{ij}$ measured in subject $i$ at the $j$th assessment time and $x_{ij} = (x_{i1j}, x_{i2j})$, with $x_{i1j} = 1$ if subject $i$ received zidovudine and 0 otherwise for all $j$, and $x_{i2j}$ being the index of time $j$ for the subject $i$. We considered the model at $\tau = 0.25$, 0.5, and 0.75 to assess the relative effect of zidovudine when compared with that of didanosine, over time in the middle of the patient population, and at the 25th and 75th percentiles. We see that $\beta_{1,\tau}$

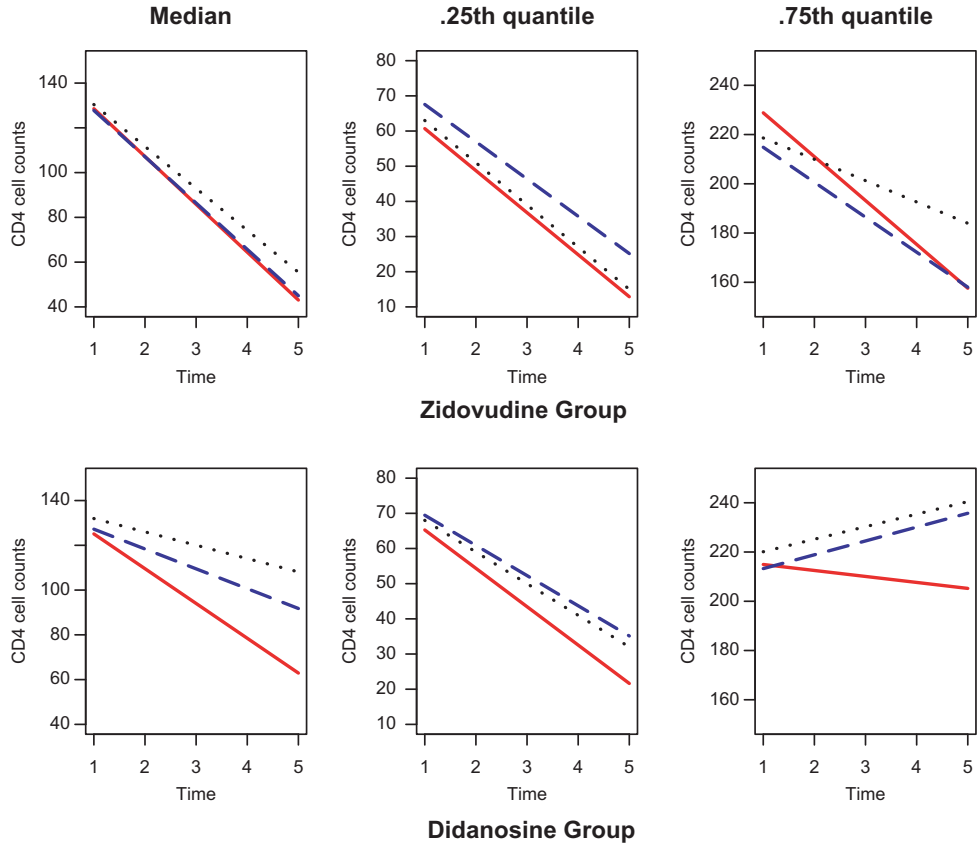


Fig. 2. Fitted CD4 cell counts based on the proposed approaches using an AR(1) working correlation under missing at random mechanism (Proposed, solid lines), an AR(1) working correlation under missing completely at random mechanism (Naive 1, dashed lines), and an independent working correlation under missing completely at random mechanism (Naive 2, dotted lines).

corresponds to the overall relative effect and $\beta_{3,\tau}$ indicates a time-related change in the relative effect, i.e. time by treatment interaction.

In the HIV study, a proportion of participants was lost to follow-up and dropout rate increased over time: in the zidovudine treatment group, 24.6% of the study subjects dropped out after the first visit and overall 60.7% dropped out after the fifth visit. In the didanosine treatment group, 15.7% of the study subjects dropped out after the first visit and overall 60.0% dropped out over the five study visits. Volberding *and others* (1990) suggested that the main reason for dropouts are the selective withdrawal from the study of subjects with low or declining CD4 cell counts. In this study, there are more patients with smaller CD4 cell counts ($<100$) for the zidovudine group than didanosine at the baseline. This suggests that the assumption of missing completely at random might not be valid in the study.

We postulated that the dropout process was dependent on the last measured CD4 cell counts, the type of treatments, and the measurement time. Following Lipsitz *and others* (1997), we applied the logistic regression to evaluate the probability of being observed at the $j$th occasion of the $i$th subject, $\text{logit}\,\lambda_{ij} = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2j} + \alpha_3 y_{i,j-1}$, where $\lambda_{ij} = P(M_{ij} = 1 \mid M_{i,j-1} = 1, x_{i1}, x_{i2}, y_{i,j-1})$ and $M_{ij} = 1$ indicates $y_{ij}$ is observed. The coefficient of the previous CD4 cell count $\alpha_3$ is statistically significant with

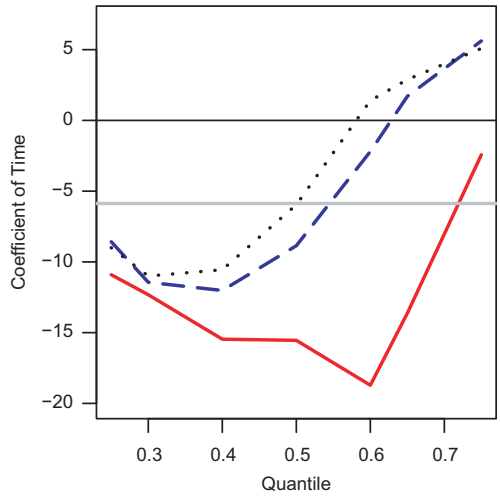


Fig. 3. Fitted quantile coefficients of a covariate "Time" with $\tau \in (0.25, 0.75)$ for the proposed approach (the solid line), Naive1 (the dashed line), and Naive2 (the dotted line). The solid horizontal gray line indicates the estimated coefficient for the mean model.

a positive estimate ($\hat{\alpha}_3 = 0.003$ with a small standard error 0.001). This indicates that patients with lower CD4 cell counts are more likely to drop out of the study. Negative estimates of $\alpha_1$ and $\alpha_2$ ($\hat{\alpha}_1 = -0.031$ and $\hat{\alpha}_2 = -0.005$) imply that a patient who received didanosine treatment is less likely to withdraw from the study, and more patients tend to be lost to follow-up as time goes on, although these trends are not statistically significant.

The equispaced measurement times and the nature of the CD4 cell counts measure suggest autocorrelative dependence among the repeatedly measured CD4 cell count data within subjects. This is translated to a toeplitz dependence structure of sign correlations for the quantile marginal regression. An AR(1) correlation structure well approximates the toeplitz structure and was used as a working correlation structure. We applied the weighted empirical likelihood procedure approach with the estimated $\lambda_{ij}$ under the missing at random assumption. As the constraint equations $\sum_{i=1}^{n} p_i \hat{\mathbf{g}}_i^w(\boldsymbol{\beta}) = 0$ are overdetermined, the maximization in (2.11) was conducted by the existing R package *optim* with the empirical likelihood (2.10) as the objective function. Given $\boldsymbol{\beta}$, the empirical likelihood is evaluated by the R package *emplik*. We first computed the standard quantile regression estimator ignoring the dependency structure and using the observed data, and used the estimate as the starting value for the maximization computation by the R package *optim*. We provide R codes used for this real-life data application as supplementary material (available at *Biostatistics* online).

We compared this approach with two naive approaches. The first naive approach used the same working correlation structure, an AR(1) structure, for the within-subject correlation, but assumed that the missingness were completely at random (Naive 1). The second naive approach also assumed completely randomness for the missing mechanism but assumed working independence for the within-subject correlation (Naive 2).

Table 3 reports the estimates of the coefficients with the standard errors estimated by bootstrapping. We evaluated $\hat{\boldsymbol{\beta}}_\tau^w$ in 1000 bootstrap samples and estimated the sample standard errors. Based on the asymptotic normality results, we constructed 95% confidence intervals and assessed the statistical significance of the coefficient estimates at the significance level of 0.05. In Table 3, statistically significant coefficients are

marked with ∗. Alternatively, we considered bootstrap percentile-based confidence intervals and obtained similar significance results.

Figures 2 and 3 show that the naive approaches tend to overestimate, when compared with the proposed approach. The bias can be explained as the naive approaches ignored the dropout process in that patients who had lower CD4 cell counts were more likely to withdraw from the study. Furthermore, Figure 3 clearly suggests negative time effects for overall quantiles by the proposed approach, while the naive approaches yielded positive estimates for higher quantiles than the median. Positive time effects not only disagree with the sample but also contradict the general knowledge that CD4 cell counts tend to decrease over time. We observe that the estimated coefficients vary across quantiles, which indicates the heterogeneity of the data. For example, the proposed approach reports that the treatment and the interaction effects were statistically significant with positive and negative estimates at the 0.75th quantile, while only the effect of time is negatively associated with the CD4 cell counts at the median and the 0.75th quantile. The mean regression model using the generalized estimating equations provides quite different results. This might be due to the fact that the distribution of the data is skewed and the missing mechanism is far from the missing completely at random mechanism. Therefore, it would not be desirable to apply the generalized estimating equations based on an assumed normal distribution to the HIV dataset.

## 4. Discussion

Various methodologies have been developed for the conditional mean analysis that readily accommodate the correlations between repeated measurements; see Huang *and others* (2006, 2007), El Karoui (2008), Fan *and others* (2008), and Zhou and Qu (2012). For the quantile marginal regression, however, most existing methodologies have overlooked the within-subject correlations with a few exceptions. Yi and He (2009) proposed incorporating the within-subject correlations for the median regression by assuming an unstructured correlation structure and jointly estimating the correlation matrix. This approach may not be generally applicable, when low or high quantiles are of interest or the number of repeated measurements is relatively large because the correlation matrix cannot be reliably estimated. Other approaches requiring the estimation of the correlation structure may be subject to similar limitations. The proposed empirical likelihood inference procedure accommodates the correlation information, while it avoids estimating the correlation matrix by using the matrix expansion idea of the quadratic information function. It yields a more efficient estimator without knowing the true correlation structure nor estimating the parameters involved in the informative working correlation structure. In addition, the proposed approach readily accounts for dropouts arising from a missing at random mechanism. With dropouts, the proposed procedure can be seen as an inverse probability weighted (IPW) estimation method. In general, IPW methods are known to be sensitive to misspecification of the probability model, particularly when some estimated probabilities are small (Kang and Schafer, 2007). In practice, incomplete longitudinal data may include non-monotone missingness. The proposed method can be expanded in such cases by transforming the unbalanced data to artificially balanced data as in Zhou and Qu (2012).

## Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

## Acknowledgments

The authors are very grateful to the Co-Editor, two referees, and an associate editor for their insightful comments and suggestions that have improved the manuscript significantly. *Conflict of Interest*: None declared.

## FUNDING

M.-O.K. research was supported by the National Science Foundation award (DMS-1007535) and H.G.H. research was supported in part by NSA grant (H98230-15-1-0260).

## REFERENCES

CHO, H. AND QU, A. (2015). Efficient estimation for longitudinal data by combining large-dimensional moment conditions. *Electronic Journal of Statistics* **9**, 1315–1334.

DOLIN, R., AMATO, D. A., FISCHL, M. A., PETTINELLI, C., BELTANGADY, M., LIOU, S., BROWN, M. J., CROSS, A. P., HIRSCH, M. S., HARDY, W. D. *and others* (1995). Zidovudine compared with didanosine in patients with advanced HIV type 1 infection and little or no previous experience with zidovudine. *Archives of Internal Medicine* **155**, 961–974.

DUNSON, D. B., WATSON, M. AND TAYLOR, J. A. (2003). Bayesian latent variable models for median regression on multiple outcomes. *Biometrics* **59**, 296–304.

EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics* **36**, 2717–2756.

FAN, J., FAN, Y. AND LV, J. (2008). High-dimensional covariance matrix estimation using a factor model. *Econometrics* **147**, 186–197.

GERACI, M. AND BOTTAI, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* **8**, 140–154.

HE, X., FU, B. AND FUNG, W. K. (2003). Median regression for longitudinal data. *Statistics in Medicine* **22**, 3655–3669.

HUANG, J. Z., LIU, L. AND LIU, N. (2007). Estimation of large covariance matrices of longitudinal data with basis function approximations. *Journal of Computational and Graphical Statistics* **16**, 189–209.

HUANG, J. Z., LIU, N., POURAHMADI, M. AND LIU, L. (2006). Covariance selection and estimation via penalised normal likelihood. *Biometrika* **93**, 85–98.

JUNG, S. (1996). Quasi-likelihood for median regression models. *Journal of the American Statistical Association* **91**, 251–257.

KANG, J. D. Y. AND SCHAFER, J. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean form incomplete data. *Statistical Science* **22**, 523–539.

KOENKER, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis* **91**, 74–89.

LENG, C. AND ZHANG, W. (2014). Smoothing combined estimating equations in quantile regression for longitudinal data. *Statistics and Computing* **1**, 123–136.

LIANG, K. Y. AND ZEGER, S. L. (1986). Longitudinal data analysis using generalised linear models. *Biometrika* **73**, 12–22.

LIPSITZ, S. R., FITZMAURICE, G. M., MOLENBERGHS, G. AND ZHAO, L. P. (1997). Quantile regression methods for longitudinal data with drop-outs: application to CD4 cell counts of patients infected with the human immunodeficiency virus. *Applied Statistics* **46**, 463–476.

LU, X. AND FAN, Z. (2015). Weighted quantile regression for longitudinal data. *Computational Statistics* **30**, 569–592.

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

QU, A., LINDSAY, B. G. AND LI, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87**, 823–836.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**, 106–121.

Tang, C. Y. and Leng, C. (2011). Empirical likelihood and quantile regression in longitudinal data analysis. *Biometrika* **98**, 1001–1006.

Volberding, P. A., Lagakos, S. W., Koch, M. A., Pettinelli, C., Myers, M. W., Booth, M. K., Balfour, H. H., Reichman, R. C., Bartlett, J. A., Hirach, M. S. *and others* (1990). Zidovudine in asymptomatic human immunodeficiency virus infection. A controlled trial in persons with fewer than 500 CD4-positive cells per cubic millimeter. *The New England Journal of Medicine* **322**, 941–949.

Wang, H. and Zhu, Z. (2011). Empirical likelihood for quantile regression models with longitudinal data. *Journal of Statistical Planning and Inference* **141**, 1603–1615.

Whang, Y. J. (2006). Smoothed empirical likelihood methods for quantile regression models. *Econometric Theory* **22**, 173–205.

Yi, G. and He, W. (2009). Median regression models for longitudinal data with dropouts. *Biometrics* **65**, 618–625.

Zhou, J. and Qu, A. (2012). Informative estimation and selection of correlation structure for longitudinal data. *Journal of the American Statistics Association* **107**, 701–710.

[*Received June 27, 2015; revised December 28, 2015; accepted for publication December 28, 2015*]