Covariate-adjusted Regression for Distorted Longitudinal Data with Informative Observation Times

Shirong Deng[a] and Xingqiu Zhao[b]

[a]School of Mathematics and Statistics, Wuhan University, Wuhan, China

[b]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong

Abstract

In many longitudinal studies, repeated response and predictors are not directly observed, but can be treated as distorted by unknown functions of a common confounding covariate. Moreover, longitudinal data involve an observation process which may be informative with a longitudinal response process in practice. To deal with such complex data, we propose a class of flexible semiparametric covariate-adjusted joint models. The new models not only allow for the longitudinal response to be correlated with observation times through latent variables and completely unspecified link functions, but they also characterize distorted longitudinal response and predictors by unknown multiplicative factors depending on time and a confounding covariate. For estimation of regression parameters in the proposed models, we develop a novel covariate-adjusted estimating equation approach which does not rely on forms of link functions and distributions of frailties. The asymptotic properties of resulting parameter estimators are established and examined by simulation studies. A longitudinal data example containing calcium adsorption and intake measurements is provided for illustration.

KEYWORDS: Asymptotic normality; Covariate-adjusted regression; Distorted longitudinal data; Informative observation times; Latent variable.

CONTACT Xingqiu Zhao (xingqiu.zhao@polyu.edu.hk), Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong

1 Introduction

In many medical studies, both response and predictor variables may not be directly observable due to influence of a confounding variable. Instead, contaminated/distorted versions of variables may be observed through a multiplicative/additive distorting factor that is an unknown smooth function of an observed confounding variable, such as body mass index (weight/height^2)(BMI) or other measures of body configuration. Such an example is the fibrinogen data in Kaysen et al. (2003), where both plasma fibrinogen concentration as response and serum transferrin level as predictor for 69 haemodialysis patients are distorted and the BMI can be taken as a confounding variable. To deal with this situation, Sent*ü*rk and M*ü*ller (2005, 2006) developed a covariate-adjusted regression (CAR) method. The CAR method has been applied to different models. For example, Cui et al. (2009) extended the covariate-adjusted regression to nonlinear model with the response and predictors distorted by multiplicative factors; Sent*ü*rk and Nguyen (2009) developed a broader class of partial covariate-adjusted regression (PCAR) models.

A typical example of distorted longitudinal response and predictors is a longitudinal data set including the calcium absorbtion and intake measurements on 188 subjects (Davis, 2002, page 336). For the analysis of such data, Sent*ü*rk (2006) proposed a covariate-adjusted varying coefficient model through body surface area (BSA) and developed a two-step estimation method. However, the asymptotic properties of the estimators have not been established.

In the longitudinal data set mentioned above, every individual was scheduled to be measured in 5-year intervals for the calcium absorption, calcium intake, age, BSA and some other variables. However, the actual measurement times were not the same as the exact

scheduled times, and the number of repeated measurements randomly ranged from 1 to 4. It is intuitive to regard these measurement times as from an underlying observation process whose jumping points are ages at measured. Here, we can take the birth year as the starting time point. As shown in Heaney et al. (1989), age had a significant influence on calcium absorption efficiency, implying that the observation process is informative with the response, calcium absorption. This example motivates us to study distorted longitudinal data with informative observation times. For the analysis of longitudinal data with informative observation times, two methods have been developed. One is the conditional modeling approach (Sun et al., 2005; Zhao et al., 2014), which directly characterized the dependence between the response process and the observation times. Another one is the frailty-based approach proposed by Sun, Sun and Liu (2007), Liang, Lu and Ying (2009), Sun, Song and Zhou (2011), Zhao, Tong and Sun (2012), and Zhou, Zhao and Sun (2013) among others.

The methods mentioned above are designed for either distorted longitudinal data with noninformative observation times or undistorted longitudinal data with informative observation times but not for distorted longitudinal data with informative observation times. To the best of our knowledge, no statistical methods can be available for analyzing distorted longitudinal data in the presence of informative measurement times in the literature. Our goal is to develop a new approach for the statistical analysis of such complex longitudinal data. For this purpose, we propose a class of flexible semiparametric covariate-adjusted joint models, where repeated responses and predictors are contaminated with unknown multiplicative functions of time and a confounder variable, and longitudinal response and observation processes are correlated through latent variables and completely unspecified link functions.

For inference, a novel covariate-adjusted estimating equation approach is developed.

The remainder of this paper is organized as follows. We begin in Section 2 by introducing notation and describing the covariate-adjusted joint models for distorted longitudinal data with informative observation times. In Section 3, a novel covariate-adjusted estimating equation approach is developed to estimate regression parameters in the proposed models. The asymptotic properties of the resulting estimators are given in Section 4. The simulation results are presented in Section 5 to assess the finite-sample performance of the proposed inference procedure. A real example of distorted longitudinal data is provided to illustrate applications of the proposed method in Section 6. Some concluding remarks are made in Section 7. All technical proofs are given in the Supplemental Materials.

2 Statistical Model

Consider a longitudinal study that consists of n independent subjects. For subject i, let $Y_i(t)$ and $\boldsymbol{X}_i(t)$ be the underlying unobserved response variable and p-dimensional vector of covariates valued at time t. We assume that $Y_i(t)$ takes a marginal model

$$E\{Y_i(t)|\boldsymbol{X}_i(t), Z_i\} = \mu_0(t) + \boldsymbol{\beta}_0'\boldsymbol{X}_i(t) + g(Z_i), \quad i = 1, \ldots, n, \tag{2.1}$$

where $\mu_0(\cdot)$ is an unknown baseline mean function, and $\boldsymbol{\beta}_0$ is a p-dimensional vector of regression coefficients, Z_i is an unobserved positive latent variable, which is independent of $\boldsymbol{X}_i(\cdot)$, and $g(\cdot)$ is a completely unspecified function with $E\{g(Z)\} = 0$ for identifiability. Model (2.1) characterizes the marginal mean of the process $Y(\cdot)$ while leaving its dependence structure and distributional form completely unspecified.

The repeated response and covariates can be observed after being contaminated by unknown functions of a common observable variable U and time. That is,

$$\widetilde{Y}_i(t) = \psi(U_i, t)Y_i(t) \quad \text{and} \quad \widetilde{X}_{ri}(t) = \phi_r(U_i, t)X_{ri}(t), \quad r = 1, \ldots, p \tag{2.2}$$

are the actual observable response and covariates valued at time t, with ψ and ϕ_r being the unknown distorting functions. The identifiability requires the condition that the distortion is mean-preserving for each t, i.e., the means of the observed variables $E\{\widetilde{Y}_i(t)\}$ and $E\{\widetilde{X}_{ri}(t)\}$ are the same as those of the underlying variables, $E\{Y_i(t)\}$ and $E\{X_{ri}(t)\}$, respectively. Under the assumption that U is independent of Y and $\boldsymbol{X}(\cdot)$, this identifiability condition is equivalent to the following constraints

$$E\{\psi(U, t)\} = 1 \quad \text{and} \quad E\{\phi_r(U, t)\} = 1, \quad r = 1, \ldots, p, \tag{2.3}$$

for time t. Models (2.1)–(2.3) will be referred to as the covariate-adjusted marginal model for distorted longitudinal data.

Suppose that $\widetilde{Y}_i(t)$ is observed at distinct time points $0 < T_{i,1} < T_{i,2} < \cdots < T_{i,K_i}$, where K_i is the potential number of observations on subject i. In the following, we regard these observation times arising from an underlying counting process $N_i^*(t)$ characterized by $N_i(t) = \sum_{k=1}^{K_i} I(T_{ik} \leq t) = N^*(\min(t, C_i))$, where $I(\cdot)$ is the indicator function, and C_i is the follow-up or censoring time with $K_i = N_i^*(C_i)$ for subject $i, i = 1, \ldots, n$. Then, the processes $\widetilde{Y}_i(t)$ is observed only at the time points where $N_i(t)$ jumps. The covariate $\{\widetilde{\boldsymbol{X}}_i(t), 0 \leq t \leq C_i, i = 1, \ldots, n\}$ are assumed to be observed.

Following Liang, Lu and Ying (2009), we assume that the potential observation process

$N_i^*(t)$ is a mixed Poisson process with the intensity function

$$\lambda(t|\boldsymbol{V}_i, Z_i) = \lambda_0(t) Z_i h(\boldsymbol{V}_i), \quad i = 1, \ldots, n, \tag{2.4}$$

where $\lambda_0(t)$ is a completely unknown baseline intensity function, $h(\cdot)$ is a completely unspecified positive function and $\boldsymbol{V}_i$ is a vector of l-dimensional baseline covariates, which is independent of the frailty Z_i. Let $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ and take $\Lambda_0(\tau) = 1$ for the identifiability with τ being the length of the study.

Remark 1. *In model (2.1), $g(\cdot)$ takes a role of an unknown link function that is used to characterize the relationship between the longitudinal response and observation processes. To see this, we suppose that $Y_i(t)$ follows a semiparametric random effects model*

$$Y_i(t) = \mu_0(t) + \boldsymbol{\beta}_0' \boldsymbol{X}_i(t) + \eta_i + \varepsilon_i(t), \quad i = 1, \ldots, n,$$

where η_i is a random variable of subject-specific effect, and $\varepsilon_i(t)$ is a zero mean measurement error process. Taking the conditional expectation of $Y_i(t)$ yields model (2.1) with $g(Z_i) = E(\eta_i|Z_i)$. For example, if one takes $g(Z) = \rho\{Z - E(Z)\}$, then ρ characterizes the relationship between the observation process and the longitudinal response process. When $\rho > 0$ ($\rho < 0$), the two processes are positively (negatively) correlated; when $\rho = 0$, the two processes have no correlation given the covariates.

For inference on models (2.1)-(2.4), we need some basic assumptions: (A1) conditional on $\boldsymbol{X}(\cdot)$, $\boldsymbol{V}$ and Z, the processes $N^*(\cdot)$ and $Y(\cdot)$ are independent; (A2) conditional on $\boldsymbol{X}(\cdot)$ and $\boldsymbol{V}$, censoring time C is independent of $N^*(\cdot)$, $Y(\cdot)$ and Z (C is noninformative); (A3) U is independent of $\boldsymbol{X}(\cdot)$, $\mathbf{V}$, $Y(\cdot)$, $N^*(\cdot)$, C, and Z.

The observed data consist of

$$\{\boldsymbol{O}_i = (K_i, \overline{T}_{iK_i}, \overline{N}_{iK_i}, U_i, \overline{\widetilde{Y}}_{iK_i}, C_i, \widetilde{\boldsymbol{X}}_i(t), \boldsymbol{V}_i), 0 \le t \le C_i, i = 1, \ldots, n\}$$

with $\overline{T}_{iK_i} = (T_{i,1}, \ldots, T_{i,K_i})'$, $\overline{N}_{iK_i} = (N_i(T_{i,1}), \ldots, N_i(T_{i,K_i}))'$, $\overline{\widetilde{Y}}_{iK_i} = (\widetilde{Y}_i(T_{i,1}), \ldots, \widetilde{Y}_i(T_{i,K_i}))'$. The central goal of this paper is to estimate coefficient $\boldsymbol{\beta}_0$ in (2.1) based on the observed data.

3 Estimation Procedure

3.1 Estimating equation for $\boldsymbol{\beta}_0$ based on uncontaminated data

In this subsection, we present an estimating equation for $\boldsymbol{\beta}_0$ when $Y_i(t)$ and $\boldsymbol{X}_i(t)$ are observable. First, it follows from Liang, Lu and Ying (2009) that

$$E\{\xi_i(t)dN_i^*(t)|\boldsymbol{V}_i, Z_i, C_i, K_i\} = \xi_i(t)K_i\Lambda_0(C_i)^{-1}d\Lambda_0(t), \tag{3.1}$$

where $\xi_i(t) = I(C_i \ge t)$. Then, under assumptions (A1) and (A2), by (3.1), we have

$$\begin{aligned}
&E\left[K_i^{-1}\{Y_i(t) - \boldsymbol{\beta}_0'\boldsymbol{X}_i(t)\}dN_i(t)|\boldsymbol{X}_i(t), \boldsymbol{V}_i, C_i\right] \\
=&E\left[E\left\{K_i^{-1}(Y_i(t) - \boldsymbol{\beta}_0'\boldsymbol{X}_i(t))dN_i(t)|\boldsymbol{X}_i(t), \boldsymbol{V}_i, C_i, K_i, Z_i\right\}|\boldsymbol{X}_i(t), \boldsymbol{V}_i, C_i\right] \\
=&E\left[K_i^{-1}E\left\{(Y_i(t) - \boldsymbol{\beta}_0'\boldsymbol{X}_i(t))|\boldsymbol{X}_i(t), C_i, K_i, Z_i\right\}E\{dN_i(t)|\boldsymbol{V}_i, C_i, K_i, Z_i\}|\boldsymbol{X}_i(t), \boldsymbol{V}_i, C_i\right] \\
=&E\left[K_i^{-1}\{\mu_0(t) + g(Z_i)\}\xi_i(t)K_i\Lambda_0(C_i)^{-1}d\Lambda_0(t)|\boldsymbol{X}_i(t), \boldsymbol{V}_i, C_i\right] \\
=&\xi_i(t)\left[\mu_0(t) + E\{g(Z_i)|\boldsymbol{X}_i(t), \boldsymbol{V}_i, C_i\}\right]\Lambda_0(C_i)^{-1}d\Lambda_0(t) \\
=&\xi_i(t)\mu_0(t)\Lambda_0(C_i)^{-1}d\Lambda_0(t) = \xi_i(t)\Lambda_0(C_i)^{-1}d\mathcal{A}_0(t),
\end{aligned} \tag{3.2}$$

where $\mathcal{A}_0(t) = \int_0^t \mu_0(s)d\Lambda_0(s)$. Define

$$dM_i(t;\boldsymbol{\beta}_0,\Lambda_0,\mathcal{A}_0,Y_i(\cdot),\boldsymbol{X}_i(\cdot)) = \xi_i(t)\left[K_i^{-1}\{Y_i(t)-\boldsymbol{\beta}_0'\boldsymbol{X}_i(t)\}dN_i^*(t) - \Lambda_0(C_i)^{-1}d\mathcal{A}_0(t)\right].$$

Then it follows from (3.2) that

$$E\{dM_i(t;\boldsymbol{\beta}_0,\Lambda_0,\mathcal{A}_0,Y_i(\cdot),\boldsymbol{X}_i(\cdot))\} = 0. \tag{3.3}$$

Thus, for given $\boldsymbol{\beta}_0$, a reasonable estimator for $\mathcal{A}_0$ is the solution to

$$\sum_{i=1}^n M_i(t;\boldsymbol{\beta}_0,\Lambda_0,\mathcal{A},Y_i(\cdot),\boldsymbol{X}_i(\cdot)) = 0, \quad 0 \le t \le \tau.$$

Denote this estimator by $\hat{\mathcal{A}}_0(t;\boldsymbol{\beta}_0,\Lambda_0,Y_i(\cdot)'s,\boldsymbol{X}_i(\cdot)'s)$, which can be expressed as

$$\hat{\mathcal{A}}_0(t;\boldsymbol{\beta}_0,\Lambda_0,Y_i(\cdot)'s,\boldsymbol{X}_i(\cdot)'s) = n^{-1}\sum_{i=1}^n \int_0^t \frac{K_i^{-1}[Y_i(u)-\boldsymbol{\beta}_0'\boldsymbol{X}_i(u)]}{S^{(0)}(u)}dN_i(u), \tag{3.4}$$

where

$$S^{(0)}(t) = n^{-1}\sum_{i=1}^n \xi_i(t)\Lambda_0(C_i)^{-1}.$$

To estimate $\boldsymbol{\beta}_0$, we construct a proper estimating function for $\boldsymbol{\beta}$. On one hand,

$$\begin{aligned}
&\sum_{i=1}^n \int_0^\tau \boldsymbol{X}_i(t)dM_i(t;\boldsymbol{\beta},\Lambda_0,\hat{\mathcal{A}}_0(t;\boldsymbol{\beta},\Lambda_0,Y_i(\cdot)'s,\mathbf{X}_i(\cdot)'s),Y_i(\cdot),\mathbf{X}_i(\cdot)) \\
=&\sum_{i=1}^n \int_0^\tau K_i^{-1}\boldsymbol{X}_i(t)\{Y_i(t)-\boldsymbol{\beta}'\boldsymbol{X}_i(t)\}dN_i(t) \\
&-\sum_{i=1}^n \int_0^\tau \xi_i(t)\boldsymbol{X}_i(t)\Lambda_0(C_i)^{-1}\frac{n^{-1}\sum_{j=1}^n K_j^{-1}[Y_j(t)-\boldsymbol{\beta}'\boldsymbol{X}_j(t)]}{S^{(0)}(t)}dN_j(t) \\
=&\sum_{i=1}^n \int_0^\tau K_i^{-1}[\boldsymbol{X}_i(t)-\overline{\boldsymbol{X}}(t)][Y_i(t)-\boldsymbol{\beta}'\boldsymbol{X}_i(t)]dN_i(t),
\end{aligned} \tag{3.5}$$

where $\overline{\boldsymbol{X}}(t) = S_X^{(1)}(t)/S^{(0)}(t)$ with

$$S_X^{(1)}(t) = n^{-1}\sum_{i=1}^n \xi_i(t)\boldsymbol{X}_i(t)\Lambda_0(C_i)^{-1}.$$

On the other hand,

$$
\begin{aligned}
& n^{-1}\sum_{i=1}^{n}\int_0^{\tau} K_i^{-1}\{\boldsymbol{X}_i(t)-\overline{\boldsymbol{X}}(t)\}\boldsymbol{\beta}_0'\overline{\boldsymbol{X}}(t)dN_i(t) \\
=& E\left[\int_0^{\tau} K_1^{-1}\left\{\boldsymbol{X}_1(t)-\overline{\boldsymbol{x}}(t)\right\}\boldsymbol{\beta}_0'\overline{\boldsymbol{x}}(t)dN_1(t)\right]+o_p(1) \\
=& E\left[\int_0^{\tau}\beta_0'\overline{\boldsymbol{x}}(t)K_1^{-1}\left\{\xi_1(t)\boldsymbol{X}_1(t)E\left(dN_1^*(t)\middle|\boldsymbol{X}_1(t),\boldsymbol{V}_1,C_1,K_1\right)\right.\right. \\
& \qquad\left.\left.-\overline{\boldsymbol{x}}(t)\xi_1(t)E\left(dN_1^*(t)\middle|\boldsymbol{X}_1(t),\boldsymbol{V}_1,C_1,K_1\right)\right\}\right]+o_p(1) \\
=& \int_0^{\tau}\beta_0'\overline{\boldsymbol{x}}(t)\left[E\left\{\xi_1(t)\boldsymbol{X}_1(t)\Lambda_0(C_1)^{-1}\right\}-\overline{\boldsymbol{x}}(t)E\left\{\xi_1(t)\Lambda_0(C_1)^{-1}\right\}\right]d\Lambda_0(t)+o_p(1) \\
=& o_p(1), \qquad (3.6)
\end{aligned}
$$

where $\overline{\boldsymbol{x}}(t)$ is the limit of $\overline{\boldsymbol{X}}(t)$. Similarly, we have

$$
n^{-1}\sum_{i=1}^{n}\int_0^{\tau} K_i^{-1}\{\boldsymbol{X}_i(t)-\overline{\boldsymbol{X}}(t)\}\overline{Y}(t)dN_i(t)=o_p(1), \qquad (3.7)
$$

where $\overline{Y}(t)=S_Y^{(1)}(t)/S^{(0)}(t)$ with

$$
S_Y^{(1)}(t)=n^{-1}\sum_{i=1}^{n}\xi_i(t)Y_i(t)\Lambda_0(C_i)^{-1}.
$$

Therefore, if Λ_0 is known, by combing (3.3), (3.5), (3.6) and (3.7), we can estimate $\boldsymbol{\beta}_0$ through the following estimating equation

$$
\begin{aligned}
& \widetilde{W}(\boldsymbol{\beta};\Lambda_0,Y_i(\cdot)'s,\boldsymbol{X}_i(\cdot)'s) \\
=& \sum_{i=1}^{n}\int_0^{\tau} K_i^{-1}\{\boldsymbol{X}_i(t)-\overline{\boldsymbol{X}}(t)\}[\{Y_i(t)-\overline{Y}(t)\}-\boldsymbol{\beta}'\{\boldsymbol{X}_i(t)-\overline{\boldsymbol{X}}(t)\}]dN_i(t)=0. \qquad (3.8)
\end{aligned}
$$

Since $\hat{\mathcal{A}}_0(t;\boldsymbol{\beta},\Lambda_0,Y_i(\cdot)'s,\boldsymbol{X}_i(\cdot)'s)$ given in (3.4) and the estimating equation (3.8) involve the unknown function Λ_0, we need to estimate it. Following Wang, Qin and Chiang (2001), we can use the nonparametric maximum likelihood estimator (NPMLE) for Λ_0 as follows.

Let $\{s_l, l = 1, \ldots, m\}$ denote the ordered and distinct values of all observation times $\{T_{i,j} : j = 1, \cdots, K_i, i = 1, \ldots, n\}$ for the longitudinal response variable. Let $q_l = \sum_{i=1}^{n} dN_i^*(s_l)$ be the number of observations at s_l, and $N_l = \sum_{i=1}^{n} I(s_l \leq C_i) N_i^*(s_l)$ be the total number of observations with observation times and censoring time satisfying $T_{i,j} \leq s_l \leq C_i$. Then we can derive the conditional likelihood function of the observed data on the $T_{i,j}$'s conditional on $\{K_i, C_i, \mathbf{V}_i, Z_i\}$, and the NPMLE $\hat{\Lambda}_0(t)$ of $\Lambda_0(t)$ can be given by

$$\hat{\Lambda}_0(t) = \prod_{s_l > t} \left(1 - \frac{q_l}{N_l}\right),$$

where the product is taken to be 1 if there is no s_l with $s_l > t$.

Thus, we have the following estimating equation for $\boldsymbol{\beta}$ by replacing Λ_0 with its estimator,

$$\begin{aligned}
&\widetilde{W}(\boldsymbol{\beta}; \hat{\Lambda}_0, Y_i(\cdot)'s, \boldsymbol{X}_i(\cdot)'s) \\
=&\sum_{i=1}^{n} \int_0^{\tau} K_i^{-1} \{\boldsymbol{X}_i(t) - \widetilde{\overline{\boldsymbol{X}}}(t)\}[\{Y_i(t) - \widetilde{\overline{Y}}(t)\} - \boldsymbol{\beta}'\{\boldsymbol{X}_i(t) - \widetilde{\overline{\boldsymbol{X}}}(t)\}] dN_i(t) = 0, \qquad (3.9)
\end{aligned}$$

where $\widetilde{\overline{\boldsymbol{X}}}(t) = \tilde{S}_X^{(1)}(t)/\hat{S}^{(0)}(t)$ and $\widetilde{\overline{Y}}(t) = \tilde{S}_Y^{(1)}(t)/\hat{S}^{(0)}(t)$ with

$$\begin{aligned}
\hat{S}^{(0)}(t) =& n^{-1} \sum_{i=1}^{n} \xi_i(t) \hat{\Lambda}_0(C_i)^{-1}, \\
\tilde{S}_X^{(1)}(t) =& n^{-1} \sum_{i=1}^{n} \xi_i(t) \boldsymbol{X}_i(t) \hat{\Lambda}_0(C_i)^{-1}, \\
\tilde{S}_Y^{(1)}(t) =& n^{-1} \sum_{i=1}^{n} \xi_i(t) Y_i(t) \hat{\Lambda}_0(C_i)^{-1}.
\end{aligned}$$

3.2 Estimating equation for $\boldsymbol{\beta}_0$ based on contaminated data

In this subsection, we aim to estimate $\boldsymbol{\beta}_0$ under models (2.1)-(2.4).

For each fixed t, we have

$$\psi(U, t) = \frac{E[\tilde{Y}(t)|U]}{E[Y(t)]}, \quad \phi_r(U, t) = \frac{E[\tilde{X}_r(t)|U]}{E[X_r(t)]}, \quad r = 1, \ldots, p. \qquad (3.10)$$

For convenience, we denote the density function of U by $p(u)$ and define

$$g_Y(U,t) = E[\tilde{Y}(t)|U]p(U) \quad \text{and} \quad g_r(U,t) = E[\tilde{X}_r(t)|U]p(U), \quad r = 1,\ldots,p. \tag{3.11}$$

Using the idea of Cui et al. (2009), based on (3.10) and (3.11), we can use the kernel estimators for $\psi(U,t)$ and $\phi_r(U,t), r = 1,\ldots,p$ as follows:

$$\begin{aligned}
\hat{\psi}(u,t) =& \frac{1/(nh)\sum_{i=1}^n K((u-U_i)/h)\tilde{Y}_i(t)}{1/(nh)\sum_{i=1}^n K((u-U_i)/h)} \times \frac{1}{\overline{\tilde{Y}}(t)} \triangleq \frac{\hat{g}_Y(u,t)}{\hat{p}(u)} \times \frac{1}{\overline{\tilde{Y}}(t)}, \\
\hat{\phi}_r(u,t) =& \frac{1/(nh)\sum_{i=1}^n K((u-U_i)/h)\tilde{X}_{ri}(t)}{1/(nh)\sum_{i=1}^n K((u-U_i)/h)} \times \frac{1}{\overline{\tilde{X}}_r(t)} \triangleq \frac{\hat{g}_r(u,t)}{\hat{p}(u)} \times \frac{1}{\overline{\tilde{X}}_r(t)},
\end{aligned} \tag{3.12}$$

where $\overline{\tilde{Y}}(t) = n^{-1}\sum_{i=1}^n \tilde{Y}_i(t), \overline{\tilde{X}}_r(t) = n^{-1}\sum_{i=1}^n \tilde{X}_{ri}(t)$, h is a bandwidth, and $K(\cdot)$ is a suitable kernel function. Let

$$\hat{Y}_i(t) = \tilde{Y}_i(t)/\hat{\psi}(U_i,t), \quad \hat{X}_{ri}(t) = \tilde{X}_{ri}(t)/\hat{\phi}_r(U_i,t) \quad \text{and} \quad \hat{\boldsymbol{X}}_i(t) = (\hat{X}_{1i}(t),\cdots,\hat{X}_{pi}(t))'. \tag{3.13}$$

Substituting $Y_i(t)$ and $\boldsymbol{X}_i(t)$ by their estimates $\hat{Y}_i(t)$ and $\hat{\boldsymbol{X}}_i(t)$ in $\widetilde{W}(\boldsymbol{\beta};\hat{\Lambda}_0, Y_i(\cdot)'s, \boldsymbol{X}_i(\cdot)'s)$ of (3.9), we obtain the final working estimating equation for $\boldsymbol{\beta}$ as follows:

$$\begin{aligned}
W(\boldsymbol{\beta}) &\triangleq W(\boldsymbol{\beta};\hat{\Lambda}_0, \hat{Y}_i(\cdot)'s, \hat{\boldsymbol{X}}_i(\cdot)'s) \\
&= \sum_{i=1}^n \int_0^\tau K_i^{-1}\{\hat{\boldsymbol{X}}_i(t) - \widehat{\overline{\boldsymbol{X}}}(t)\}[\hat{Y}_i(t) - \widehat{\overline{Y}}(t) - \boldsymbol{\beta}'\{\hat{\boldsymbol{X}}_i(t) - \widehat{\overline{\boldsymbol{X}}}(t)\}]dN_i(t) = 0,
\end{aligned} \tag{3.14}$$

where $\widehat{\overline{\boldsymbol{X}}}(t) = \hat{S}_X^{(1)}(t)/\hat{S}^{(0)}(t)$ and $\widehat{\overline{Y}}(t) = \hat{S}_Y^{(1)}(t)/\hat{S}^{(0)}(t)$, with

$$\hat{S}_X^{(1)}(t) = n^{-1}\sum_{i=1}^n \xi_i(t)\hat{\boldsymbol{X}}_i(t)\hat{\Lambda}_0(C_i)^{-1} \quad \text{and} \quad \hat{S}_Y^{(1)}(t) = n^{-1}\sum_{i=1}^n \xi_i(t)\hat{Y}_i(t)\hat{\Lambda}_0(C_i)^{-1}.$$

Solving the above equation (3.14), the estimator for $\boldsymbol{\beta}_0$ has a closed form

$$\begin{aligned}
\hat{\boldsymbol{\beta}} =& \left\{\sum_{i=1}^n \int_0^\tau K_i^{-1}\{\hat{\boldsymbol{X}}_i(t) - \widehat{\overline{\boldsymbol{X}}}(t)\}^{\otimes 2}dN_i(t)\right\}^{-1} \\
&\times \left\{\sum_{i=1}^n \int_0^\tau K_i^{-1}\{\hat{Y}_i(t) - \widehat{\overline{Y}}(t)\}\{\hat{\boldsymbol{X}}_i(t) - \widehat{\overline{\boldsymbol{X}}}(t)\}dN_i(t)\right\},
\end{aligned}$$

where $a^{\otimes 2} = aa'$ for a column vector a. Then, $\mathcal{A}_0$ can be estimated by

$$\hat{\mathcal{A}}_0(t) = \hat{\mathcal{A}}_0(t; \hat{\boldsymbol{\beta}}, \hat{\Lambda}_0, \hat{Y}_i(\cdot)'s, \hat{\boldsymbol{X}}_i(\cdot)'s) = n^{-1} \sum_{i=1}^{n} \int_0^t \frac{K_i^{-1}[\hat{Y}_i(u) - \hat{\beta}'\hat{\boldsymbol{X}}_i(u)]}{\hat{S}^{(0)}(u)} dN_i(u).$$

4 Asymptotic Properties

To establish the asymptotic properties of the estimators, we need the following regularity conditions.

C1. $P(C \geq \tau, Z > 0) > 0$, $E(Z^2) < \infty$, and $P(C > \tau_\eta) = 1$, where $\tau_\eta = \inf\{t : \Lambda_0(t) > \eta\}$ for some $\eta > 0$.

C2. $\boldsymbol{X}_i(t), Y_i(t), i = 1, \cdots, n$ have bounded total variations, i.e. $|X_{ri}(0)| + \int_0^\tau |dX_{ri}(t)| \leq M_0$ and $|Y_i(0)| + \int_0^\tau |dY_i(t)| \leq M_1$ for all $r = 1, \cdots, p$ and $i = 1, \cdots, n$, where X_{ri} is the rth component of $\boldsymbol{X}_i$ and M_0, M_1 are constants.

C3. $E\{N_i^*(\tau)\} < \infty$ and $K_i \geq 1$ $(i = 1, \ldots, n)$.

C4. For each t, $g_r(u;t) = E[X_r(t)]\phi_r(u,t)p(u), 1 \leq r \leq p$, $g_Y(u,t) = E[Y(t)]\psi(u,t)p(u)$ and $p(u)$ are 3-times differential with respect to u, and their third derivatives satisfy the condition that there exists a neighborhood of the origin, say, $\triangle$ and a constant $c > 0$ such that, for any $\delta \in \triangle$,

$$\begin{aligned} |g_r^{(3)}(u+\delta, t) - g_r^{(3)}(u,t)| \leq & c\delta, \quad 1 \leq r \leq p, \\ |g_Y^{(3)}(u+\delta, t) - g_Y^{(3)}(u,t)| \leq & c\delta, \\ |p^{(3)}(u+\delta) - p^{(3)}(u)| \leq & c\delta. \end{aligned}$$

Furthermore, $|g_r(u,t)|, 1 \leq r \leq p$, $|g_Y(u,t)|$ and $p(u)$ are greater than a positive constant and less than another positive constant, $\phi_r(u,t)$ and $\psi(u,t)$ are bounded.

C5. The continuous kernel function $K(\cdot)$ satisfies the following properties:

(a1) the support of $K(\cdot)$ is the interval $[-1,1]$;

(a2) $K(\cdot)$ is symmetric about zero;

(a3) $\int_{-1}^{1} K(u)du = 1$, $\int_{-1}^{1} u^i K(u)du = 0, i = 1,2,3$.

C6. The bandwidth h is in the range from $O(n^{-\frac{1}{4}}\log n)$ to $O(n^{-\frac{1}{8}})$.

C7. $|EY(t)|$ and $|EX_r(t)|$ are bounded away from zero.

C8. $\Gamma \triangleq E\left\{\int_0^\tau K^{-1}\{\boldsymbol{X}(t) - \bar{\boldsymbol{x}}(t)\}^{\otimes 2} dN(t)\right\}$ is positive definite.

These are all mild conditions that could be satisfied in usual situations. The boundedness conditions in C2 and C3 simplify the derivation of the asymptotic results. C4 is the boundedness and smoothness condition for functions $g_r(\cdot;t)$ and $g_Y(\cdot;t)$ and the density function $p(\cdot)$ of U. C5 and C6 are commonly used for the asymptotic properties of kernel based estimation (see, e.g., Zhu and Fang, 1996 and Cui et al., 2009). C8 can be interpreted that the sample covariance is asymptotically nonsingular.

To present the asymptotic normality for $\hat{\boldsymbol{\beta}}$, we define

$$\hat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau K_i^{-1}\{\hat{\boldsymbol{X}}_i(t) - \widehat{\bar{\boldsymbol{X}}}(t)\}^{\otimes 2} dN_i(t),$$

and

$$\hat{\boldsymbol{w}}_{1i} = \int_0^\tau \{\hat{\boldsymbol{X}}_i(t) - \widehat{\bar{\boldsymbol{X}}}(t)\} d\hat{M}_i(t; \hat{\boldsymbol{\beta}}, \hat{\Lambda}_0, \hat{\mathcal{A}}_0, \hat{Y}_i(\cdot), \hat{\boldsymbol{X}}_i(\cdot)),$$

$$\hat{\boldsymbol{w}}_{2i} = \frac{1}{n} \sum_{j=1}^n \int_0^\tau K_j^{-1} \hat{Y}_j(t) \{\hat{\boldsymbol{X}}_j(t) - \widehat{\bar{\boldsymbol{X}}}(t)\} \frac{\tilde{Y}_i(t) - \hat{Y}_i(t) + \{\hat{Y}_i(t) - n^{-1} \sum_{l=1}^n \tilde{Y}_l(t)\}/2}{n^{-1} \sum_{l=1}^n \tilde{Y}_l(t)} dN_j(t),$$

$$\hat{\boldsymbol{w}}_{3i} = - \left[\frac{1}{n} \sum_{j=1}^n \int_0^\tau \widehat{\boldsymbol{H}}_i(t) K_j^{-1} \hat{\boldsymbol{X}}_j(t) \{\hat{\boldsymbol{X}}_j(t) - \widehat{\bar{\boldsymbol{X}}}(t)\}' dN_j(t) \right]' \hat{\boldsymbol{\beta}},$$

$$\hat{\boldsymbol{w}}_{4i} = \frac{1}{n} \sum_{j=1}^n \int_0^\tau \widehat{\boldsymbol{H}}_i(t) K_j^{-1} \hat{\boldsymbol{X}}_j(t) \{\hat{Y}_j(t) - \widehat{\bar{Y}}(t) - \hat{\boldsymbol{\beta}}'(\hat{\boldsymbol{X}}_j(t) - \widehat{\bar{\boldsymbol{X}}}(t))\} dN_j(t),$$

where

$$\hat{M}_i(t; \hat{\boldsymbol{\beta}}, \hat{\Lambda}_0, \hat{\mathcal{A}}_0, \hat{Y}_i(\cdot), \hat{\boldsymbol{X}}_i(\cdot)) = \int_0^t \xi_i(u) \left[K_i^{-1} \{\hat{Y}_i(u) - \hat{\boldsymbol{\beta}}' \hat{\boldsymbol{X}}_i(u)\} dN_i^*(u) - \hat{\Lambda}_0(C_i)^{-1} d\hat{\mathcal{A}}_0(u) \right],$$

and $\widehat{\boldsymbol{H}}_i(t)$ is the diagonal matrix with its r-th diagonal element being

$$\hat{H}_{rri}(t) = \frac{\{\tilde{X}_{ri}(t) - \hat{X}_{ri}(t)\} + \{\hat{X}_{ri}(t) - n^{-1} \sum_{l=1}^n \tilde{X}_{rl}(t)\}/2}{n^{-1} \sum_{l=1}^n \tilde{X}_{rl}(t)}.$$

The asymptotic normality for $\hat{\boldsymbol{\beta}}$ is summarized as follows.

Theorem 4.1 *Under conditions C1 - C8, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ converges in distribution to a random normal variable with mean zero and a covariance matrix $\boldsymbol{\Sigma} = E(\boldsymbol{a}_1^{\otimes 2})$, which can be consistently estimated by $\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^n \hat{\boldsymbol{a}}_i^{\otimes 2}$, where $\hat{\boldsymbol{a}}_i = \hat{\Gamma}^{-1} \hat{\boldsymbol{w}}_i$ with $\hat{\boldsymbol{w}}_i = \sum_{j=1}^4 \hat{\boldsymbol{w}}_{ji}$.*

By Theorem 4.1, an approximate $(1 - \alpha)$ asymptotic confidence interval for β_r is

$$\left[\hat{\beta}_r - z_{\alpha/2} \sqrt{\frac{\hat{\sigma}_{rr}}{n}}, \hat{\beta}_r + z_{\alpha/2} \sqrt{\frac{\hat{\sigma}_{rr}}{n}} \right],$$

where $z_{\alpha/2}$ is the $(1 - \alpha/2)$th quantile of the standard normal distribution, and $\hat{\sigma}_{rr}$ is the rth diagonal element of $\hat{\boldsymbol{\Sigma}}$.

5 Simulation Studies

In this section, we conducted Monte Carlo simulation studies to evaluate the finite sample properties of the proposed estimators. We generated the latent variable Z_i from a gamma distribution with mean 1 and variance 0.25. For the longitudinal response process, we generated them from the following model:

$$Y_i(t) = \mu_0(t) + \beta_0 X_i(t) + g(Z_i) + \varepsilon_i(t),$$

where $\mu_0(t) = \log(1+t)$, and $g(Z_i) = \rho(Z_i - 1)$ with $\rho = -0.5, 0$, and 0.5, and $\varepsilon_i(t)$'s are independent standard normal variables. For the covariate $X_i(t)$, we considered the following two situations:

(a1) The time-independent covariate X_i follows the uniform distribution over interval $(0, 1)$;

(a2) The time-dependent covariate $X_i(t)$ takes the form $w_i \log(t)$, where w_i has a uniform distribution over interval $(0.5, 1)$.

Suppose that $Y_i(t)$ and $X_i(t)$ are distorted by (2.2). The confounding covariate U was simulated from $Unif(0.5, 1.5)$; the distorting functions were chosen as $\psi(U;t) = \frac{12(U+t)^2}{13+24t+12t^2}$ and $\phi(U) = \frac{U+1}{2}$ for case (a1), and $\psi(U,t) = \frac{12(U+t)^2}{13+24t+12t^2}$ and $\phi(U,t) = \frac{U+t}{1+t}$ for case (a2).

For generation of censoring time C_i, we considered two cases as follows:

(c1) (*Covariate-independent case*) C_i follows a uniform distribution over interval $(\tau/2, \tau)$;

(c2) (*Covariate-dependent case*) $C_i = \min\{C_i^*, \tau\}$, where C_i^* satisfies $\log(C_i^*) = 4(1+2V_i) + e_i$ with V_i being the same as X_i in (a1) and $e_i \sim N(0, 1)$.

In both cases, $\tau = 18$.

For the generation of the observation process $N_i^*(t)$, we considered a mixed homogeneous Poisson process with $\lambda_0(t) = 1/\tau$, that is, given V_i, C_i, and Z_i, K_i was generated from the Poisson distribution with mean $2Z_iC_i \exp\{V_i\}/\tau$. Given K_i, the observation times $(T_{i,1}, \cdots, T_{i,K_i})$ were taken to be the order statistics of the random sample of size K_i from the uniform distribution over $(0, C_i)$.

We took the true value for β_0 as -1, 0 and 1, representing different effects of the covariates $X(t)$ on the longitudinal response variable. The higher order kernel function $K(t) = \frac{15}{32}(3 - 7t^2)(1 - t^2)I(|t| \leq 1)$ which satisfies $\int_{-1}^{1} K(u)du = 1$ and $\int_{-1}^{1} u^i K(u)du = 0, i = 1, 2, 3$ was used for estimation. For the bandwidth selection, we chose interval $(0.05, 1.5)$ as the range of h to satisfy condition C6, and partitioned this interval into a grid of values:

$$h_{min} = h_0 = 0.05 < h_1 = h_0 + \delta < h_2 = h_0 + 2\delta < \cdots < h_{20} = h_{max} = 1.5$$

with $\delta = (1.5 - 0.05)/20$. Then the optimal value h^* was selected by minimizing the criterion given in (5.1) below over a grid of values $\{h_j, j = 0, 1, \ldots, 20\}$:

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau K_i^{-1}[\hat{Y}_{(-i)}(t) - \widehat{\overline{Y}}_{(-i)}(t) - \hat{\beta}_{(-i)}\{\hat{X}_{(-i)}(t) - \widehat{\overline{X}}_{(-i)}(t)\}]^2 dN_i(t), \tag{5.1}$$

where the criterion is based on the leave-one-out cross validation (Stone, 1974), $\hat{Y}_{(-i)}(t)$, $\widehat{\overline{Y}}_{(-i)}(t)$, $\hat{X}_{(-i)}(t)$, $\widehat{\overline{X}}_{(-i)}(t)$, and $\hat{\beta}_{(-i)}$ are the corresponding estimates from the data deleting the information of the i-th subject.

Tables 1 and 2 report the simulation results on estimation of β_0 for the time-dependent and time-independent covariate situations with the covariate-independent censoring time; Tables 3 and 4 display the simulation results for the time-dependent and time-independent

covariate situations with the covariate-dependent censoring time. In the tables, we compared the proposed estimation method with a competing estimation method developed by disregarding the distortion fact in the underlying response and covariates and mistaking the observed response and covariates as the underlying unobservable response and covariates in the estimating equation. The estimates for the proposed method and the competing method are given in the first row and the second row of the tables, respectively. The tables include the estimated bias (BIAS) given by the average of the estimates minus the true value, the estimated standard errors of the estimates (ESE), the sample standard deviation of the estimates (SSE), and the estimated 95% coverage probabilities (CP) obtained from 1000 independent runs.

Based on our simulation results, we have the following findings: (i) The proposed estimates are nearly unbiased in all situations considered. However, the competing estimates are obviously biased and have large biases in most situations. These facts indicate that the estimation method ignoring the distortion fact may yield an estimate with a large bias. Clearly, large biases in estimation will appear if the confounding covariate is ignored and the response and predictors are not adjusted. (ii) The sample standard errors and the estimated standard errors of the proposed estimate are close to each other. Also, the estimated 95% coverage rates are close to the nominal level, that is, the proposed procedure provides reasonable estimates and the normal approximation seems to be appropriate.

6 Application

We are interested in discovering the relationship between the calcium absorption and the calcium intake to address the problem of calcium deficiency. Heaney et al. (1989) showed that the calcium absorption was approximately inversely proportional to the square root of the calcium intake, and age had a significant influence on calcium absorption efficiency. Heaney (2003) found that the relationship between the calcium absorption and calcium intake was affected by the body configuration measures such as body mass index or body surface area (BSA), which was used as a common confounder by Sentürk (2006).

We applied the proposed method to the data analysis from a longitudinal study on 188 subjects. The aim of the study is to find out the related covariates to the calcium absorption (Davis, 2002, page 336). All the individuals were in the age ranging from 35 to 45 year at the beginning of the study (1967). Repeated measurements per individual were obtained in 5-year intervals, with the number of the repeated measurements randomly ranging from 1 to 4. The information including calcium absorption, calcium intake, age, BSA and some others was recorded at each measurement time. However, the calcium absorption and calcium intake were contaminated. In order to uncover the relationship between the underlying calcium absorption and underlying calcium intake, we assumed that calcium absorption and calcium intake can be adjusted by the common confounder BSA. For the analysis, we proposed the underlying marginal mean model for calcium absorption as follows:

$$E\{Y_i(t_{ij})|X_i(t_{ij}), Z_i\} = \mu_0(t_{ij}) + \beta_0 X_i(t_{ij}) + g(Z_i), j = 1, \ldots, K_i, i = 1, \ldots, n, \qquad (6.1)$$

where $Y_i(t_{ij})$ and $X_i(t_{ij})$ are the underlying unobservable calcium absorption (g/day) and

$1/\sqrt{\text{intake}}$ (g/day) for individual i at time point t_{ij} with t_{ij} being the age ($year$) for individual i at the j-th measurement time and K_i being the number of measurements for individual i, $\mu_0(t)$ is the baseline mean function, and Z_i is an unobservable frailty. Let $\tilde{Y}_i(t_{ij})$ and $\tilde{X}_i(t_{ij})$ denote the observable distorted calcium absorption and $1/\sqrt{\text{intake}}$, respectively, and $U_i(t_{ij})$ as the observed confounding covariate, BSA, at time point t_{ij}. In addition, we use $U_i = K_i^{-1}\sum_{l=1}^{K_i} U_i(t_{il})$ for calculation. Let $N_i(\cdot)$ represent the accumulated measurement numbers of individual i over the study period, which is assumed to follow model (2.4) with V_i being another age-independent covariate, and we took the age for individual i at the last measurement time as C_i in the analysis. The main goal here is to estimate regression coefficient β_0 based on the observed data by using the proposed method in Section 3.2.

According to Sentürk (2006), three outliers were deleted before analysis. To use our proposed estimation procedure, the kernel and bandwidth selection are the same as those in Section 5, with the selected bandwidth being $h^* = 0.63$. We obtained the BSA-adjusted estimate for calcium absorption and $1/\sqrt{\text{intake}}$, denoted as $\hat{Y}_i(t_{ij})$ and $\hat{X}_i(t_{ij})$ ($j = 1, \ldots, K_i, i = 1, \ldots, n$). Then the estimate of β_0 is obtained as 0.1908 with the standard deviation being 0.0322. Accordingly, we gave the estimate of β_0 without considering the confounding covariate BSA as 0.1890, smaller than the estimate adjusted by BSA. Figures 1 and 2 display the observed and adjusted (estimated) calcium absorption and intake. From these figures, it can be found out that the calcium intake and especially the calcium absorption are adjusted at most points. Let $s_1 < \cdots < s_m$ be ordered distinct observation times of t_{ij} 's and define the average of estimated distorting functions at each s_k as $\hat{\psi}_U(s_k) = \sum_{i=1}^{n}\sum_{j=1}^{K_i}\hat{\psi}(U_i, t_{ij})I(t_{ij} = s_k)/\sum_{i=1}^{n}\sum_{j=1}^{K_i} I(t_{ij} = s_k)$ and $\hat{\phi}_U(s_k) = \sum_{i=1}^{n}\sum_{j=1}^{K_i}\hat{\phi}(U_i, t_{ij})I(t_{ij} = s_k)/\sum_{i=1}^{n}\sum_{j=1}^{K_i} I(t_{ij} =$

s_k). Similarly, define the average of estimated distorting functions at each U_i as $\hat{\psi}_T(U_i) = K_i^{-1}\sum_{j=1}^{K_i}\hat{\psi}(U_i, t_{ij})$ and $\hat{\phi}_T(U_i) = K_i^{-1}\sum_{j=1}^{K_i}\hat{\phi}(U_i, t_{ij})$. Figure 3 shows the scatter plots with the loess (Cleveland, 1979) fitted curves (blue dotted line) for the estimated distorting functions $\hat{\psi}_U$, $\hat{\phi}_U$, $\hat{\psi}_T$ and $\hat{\phi}_T$, whose minimum and maximum values are deleted. From these figures, it can be seen that the estimated values at each s_k or U_i are all generally around 1, which just matches with the identifiability condition (2.3). Specifically, the trends for the estimated distorting functions $\hat{\psi}_U$ and $\hat{\phi}_U$ are both decreasing before age 45 and then increasing after age 45. This finding is consistent with the conclusion given in Sent*ü*rk (2006) where two groups of data observed at ages before and after 45 were analyzed separately and independently. As shown in Figure 3, the estimated distorting functions $\hat{\psi}_T$ and $\hat{\phi}_T$ are increasing and decreasing with the values of BSA, respectively.

7 Concluding Remarks

Taking into account that both the distorted response and predictors and informative observation times may exist at the same time for longitudinal data, a class of flexible semiparametric covariate-adjusted marginal joint models has been proposed. Here the longitudinal response process and the observation times are correlated through latent variables and completely unspecified link functions, and the repeated response and predictors are distorted by unknown multiplicative functions of a common confounding covariate and time. It seems that we are the first to use bivariate distorting functions of confounding covariate and time for analyzing distorted longitudinal data. The model flexibility and complexity result in more challenges for estimation, computation, and theoretical proofs. A novel covariate-adjusted estimating

equation method has been developed by using the estimators of the unobservable response and predictors obtained through nonparametric kernel estimators of the distorting functions. The estimation procedure does not rely on the forms of link functions and distributions of frailties, and thus it is robust. The asymptotic properties of the resulting estimators for the regression parameters have been established. As demonstrated in our simulation and real data studies, the proposed approaches are reasonable and applicable.

If g in model (2.1) is simpler, for example linear, that is, model (2.1) becomes

$$E\{Y_i(t)|\boldsymbol{X}_i(t), Z_i\} = \mu_0(t) + \boldsymbol{\beta}_0'\boldsymbol{X}_i(t) + \alpha_0 Z_i, \quad i = 1, \ldots, n$$

(e.g., Sun, Sun and Liu, 2007; Sun, Song and Zhou, 2011). In order to estimate α_0, we need to estimate Z_i. For this case, we take $h(\mathbf{V}) = \exp(\boldsymbol{\gamma}_0'\mathbf{V})$ in model (2.4), where $\boldsymbol{\gamma}_0$ is unknown. Furthermore, $Y_i(t)$ and $\boldsymbol{X}_i(t)$ are assumed to be distorted as (2.2) and (2.3). Let $\hat{\boldsymbol{\gamma}}$ and $\hat{\Lambda}_0$ be the consistent estimators of $\boldsymbol{\gamma}_0$ and Λ_0 using the method proposed by Huang, Qin and Wang (2010), and let $\hat{Y}_i(t)$ and $\hat{\boldsymbol{X}}_i(t)$ be the same as in (3.13). Then, to estimate $\boldsymbol{\beta}_0$ and α_0, motivated by Sun, Song and Zhou (2011), we propose using the following estimating equations:

$$U_1(\boldsymbol{\beta}, \alpha) = \sum_{i=1}^{n} \int_0^{\tau} \xi_i(t)\{\hat{\boldsymbol{X}}_i(t) - \widehat{\bar{\boldsymbol{X}}}(t)\}\{\hat{Y}_i(t) - \boldsymbol{\beta}'\hat{\boldsymbol{X}}_i(t) - \alpha\hat{Z}_i\}dN_i^*(t)$$

and

$$U_2(\boldsymbol{\beta}, \alpha) = \sum_{i=1}^{n} \int_0^{\tau} \xi_i(t)[\{\hat{Z}_i - \hat{\bar{Z}}(t)\}\{\hat{Y}_i(t) - \boldsymbol{\beta}'\hat{\boldsymbol{X}}_i(t)\} - \alpha\{\hat{\Omega}_i - \hat{Z}_i\hat{\bar{Z}}(t)\}]dN_i^*(t),$$

where $\hat{Z}_i = (K_i - 1)/\{\exp(\hat{\boldsymbol{\gamma}}'\mathbf{V}_i)\hat{\Lambda}_0(C_i)\}$, $\hat{\Omega}_i = (K_i - 1)(K_i - 2)/\{\exp(\hat{\boldsymbol{\gamma}}'\mathbf{V}_i)\hat{\Lambda}_0(C_i)\}^2$,

$$\widehat{\bar{\boldsymbol{X}}}(t) = \hat{S}_X^{(1)}(t)/\hat{S}^{(0)}(t) \quad \text{and} \quad \hat{\bar{Z}}(t) = \hat{S}_Z^{(1)}(t)/\hat{S}^{(0)}(t),$$

with

$$\hat{S}^{(0)}(t) = n^{-1}\sum_{i=1}^{n}\xi_i(t)K_i\hat{\Lambda}_0(C_i)^{-1},$$

$$\hat{S}_X^{(1)}(t) = n^{-1}\sum_{i=1}^{n}\xi_i(t)K_i\hat{\Lambda}_0(C_i)^{-1}\hat{\boldsymbol{X}}_i(t),$$

and

$$\hat{S}_Z^{(1)}(t) = n^{-1}\sum_{i=1}^{n}\xi_i(t)K_i\hat{\Lambda}_0(C_i)^{-1}\hat{Z}_i.$$

Let $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha}$ denote the solutions to $U_1(\boldsymbol{\beta},\alpha) = 0$ and $U_2(\boldsymbol{\beta},\alpha) = 0$. The asymptotic properties of $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha}$ can be established using similar arguments as used in the proof of Theorem 4.1. One can see that the resulting estimating equations involving the estimation of the latent variable and the observation process model for this case are more complicated than those for the general case considered. Since the format of the relationship between the longitudinal response and observation processes is generally unknown in practice and could be very complicated, thus a flexible model may be preferred.

Motivated by Sun et al. (2012), we can extend model (2.1) as follows:

$$E\{Y_i(t)|\boldsymbol{X}_i(t), Z_i\} = \mu_0(t; Z_i) + \boldsymbol{\beta}_0'\boldsymbol{X}_i(t), \quad i = 1,\ldots,n,$$

where $\mu_0(t, Z)$ is a completely unspecified function of t and Z including the additive and multiplicative forms of the baseline function $\mu_0(t)$ and frailty Z as special cases. Then, similar to the deduction of equation (3.2), under assumptions (A1) and (A2), we have

$$E\left\{K_i^{-1}[Y_i(t) - \boldsymbol{\beta}_0'\boldsymbol{X}_i(t)]dN_i(t)|\boldsymbol{X}_i(t), \boldsymbol{V}_i, C_i\right\} = \xi_i(t)\Lambda_0(C_i)^{-1}d\mathcal{A}_0(t),$$

where $\mathcal{A}_0(t) = \int_0^t E\{\mu_0(s, Z)\}d\Lambda_0(s)$. Thus, the same estimating equation method in Section 3 can be used.

Note that a mixed Poisson process model was assumed for the potential observation process. Such Poisson model assumption can be relaxed. Instead, the observation process can take the following rate model:

$$E\{dN_i^*(t)|\mathbf{V}_i, Z_i\} = Z_i \exp(\boldsymbol{\gamma}_0'\mathbf{V}_i)d\Lambda_0(t).$$

Under this rate model and model (2.1) with assumptions (A1) and (A2), we have

$$E\left\{[Y_i(t) - \boldsymbol{\beta}_0'\boldsymbol{X}_i(t)]dN_i(t)|\boldsymbol{X}_i(t), \boldsymbol{V}_i, C_i\right\} = \xi_i(t)\exp(\boldsymbol{\gamma}_0'\mathbf{V}_i)d\mathcal{A}_0(t),$$

where $\mathcal{A}_0(t) = \int_0^t[\mu_0(s)E(Z) + E\{Zg(Z)\}]d\Lambda_0(s)$. Let $\hat{\boldsymbol{\gamma}}$ be the consistent estimator of $\boldsymbol{\gamma}_0$ using the method proposed by Lin et al. (2000), and let $\hat{Y}_i(t)$ and $\hat{\boldsymbol{X}}_i(t)$ be the same as defined in (3.13). Then, to estimate $\boldsymbol{\beta}_0$, we propose using the following estimating equation:

$$\sum_{i=1}^{n}\int_0^{\tau}\{\hat{\boldsymbol{X}}_i(t) - \widehat{\overline{\boldsymbol{X}}}(t)\}[\hat{Y}_i(t) - \widehat{\overline{Y}}(t) - \boldsymbol{\beta}'\{\hat{\boldsymbol{X}}_i(t) - \widehat{\overline{\boldsymbol{X}}}(t)\}]dN_i(t) = 0,$$

where $\widehat{\overline{\boldsymbol{X}}}(t) = \hat{S}_X^{(1)}(t)/\hat{S}^{(0)}(t)$ and $\widehat{\overline{Y}}(t) = \hat{S}_Y^{(1)}(t)/\hat{S}^{(0)}(t)$ with

$$\hat{S}^{(0)}(t) = n^{-1}\sum_{i=1}^{n}\xi_i(t)\exp(\hat{\boldsymbol{\gamma}}'\mathbf{V}_i),$$

$$\hat{S}_X^{(1)}(t) = n^{-1}\sum_{i=1}^{n}\xi_i(t)\hat{\boldsymbol{X}}_i(t)\exp(\hat{\boldsymbol{\gamma}}'\mathbf{V}_i),$$

and

$$\hat{S}_Y^{(1)}(t) = n^{-1}\sum_{i=1}^{n}\xi_i(t)\hat{Y}_i(t)\exp(\hat{\boldsymbol{\gamma}}'\mathbf{V}_i).$$

Similarly, we can establish the corresponding asymptotic properties of the estimators.

Further research is to extend the proposed methods to other useful models such as covariate-adjusted varying coefficient regression models and covariate-adjusted partly nonlinear regression models for distorted longitudinal data. Another direction is to study covariate-adjusted regression for distorted longitudinal data with a terminal event (Sun et al., 2012;

Kong et al., 2018). Furthermore, the proposed method can be extended to the case where the confounding covariate U can be time-dependent for future research.

Supplementary Materials

The supplementary materials include the proofs of Lemmas and Theorem 4.1.

Acknowledgments

The authors would like to thank the Associate Editor and the two reviewers for their constructive and insightful comments and suggestions that greatly improved the paper. Deng's research is partly supported by the National Natural Science Foundation of China (No. 11401443, 11471252). Zhao's research is partly supported by the National Natural Science Foundation of China (No. 11771366) and The Hong Kong Polytechnic University.

References

Cui, X., Guo, W., Lin, L., and Zhu, L. (2009), "Covariate-adjusted nonlinear regression," *The Annals of Statistics,* 37, 1839–1870.

Cleveland, W. S. (1979), "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association,* 74, 829–836.

Davis, C. S. (2002), *Statistical Methods for the Analysis of Repeated Measurements*, New York: Springer.

Heaney, R. P. (2003), "Normalizing calcium intake: projected population effects for body weight," *Journal of Nutrition*, 133, 268S–270S.

Heaney, R. P., Recker, R. R., Stegman, M. R., and Moy, A. J. (1989), "Calcium absorption in women: relationships to calcium intake, estrogen status, age," *Journal of Bone and Mineral Research*, 4, 469–475.

Huang, C. Y., Qin, J., and Wang, M. C. (2010), "Semiparametric analysis for recurrent event data with time-dependent covariates and informative censoring," *Biometrics*, 66, 39–49.

Kaysen, G. A., Dubin, J. A., Muller , H. G., Mitch, W. E., Rosalses, L. M., Levin, N. W., and the HEMO study group (2003), "Relationship among inflammation nutrition and physiologic mechanisms establishing albumin levels in hemodialysis patients," *Kidney International*, 61, 2240–2249.

Kong, S., Nan, B., Kalbfleisch, J. D., Saran, R., and Hirth, R. (2018), "Conditional modeling of longitudinal data with terminal event," *Journal of the American Statistical Association*, 113, 357–368.

Liang, Y., Lu, W., and Ying, Z. (2009), "Joint modeling and analysis of longitudinal data with informative observation times," *Biometrics,* 65, 377–384.

Lin, D. Y., Wei, L. J., Yang, I., and Ying, Z. (2000), "Semiparametric regression for the mean and rate functions of recurrent events," *Journal of the Royal Statistical Society*, Series B, 62, 711–730.

Sentürk, D. (2006), "Covariate-adjusted varying coefficient models," *Biostatistics,* 7, 235–251.

Sentürk, D., and Müller, H. G. (2005), "Covariate-adjusted regression," *Biometrika*, 92, 75–89.

Sentürk, D., and Müller, H. G. (2006), "Inference for covariate adjusted regression via varying coefficient models," *The Annals of Statistics*, 34, 654–679.

Sentürk, D., and Nguyen, D. V. (2009), "Partial covariate adjusted regression," *Journal of Statistical Planning and Inference,* 139, 454–468.

Stone, M. (1974), "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society*, Series B, 36, 111–147.

Sun, J., Park, D., Sun, L., and Zhao, X. (2005), "Semiparametric regression analysis of longitudinal data with informative observation times," *Journal of the American Statistical Association*, 100, 882–889.

Sun, J., Sun, L., and Liu, D. (2007), "Regression analysis of longitudinal data in the presence of informative observation and censoring times," *Journal of the American Statistical Association,* 102, 1397–1406.

Sun, L., Song, X., and Zhou, J. (2011), "Regression analysis of longitudinal data with time-dependent covariates in the presence of informative observation and censoring times," *Journal of Statistical Planning and Inference,* 141, 2902–2919.

Sun, L., Song, X., Zhou, J., and Liu, L. (2012), "Joint analysis of longitudinal data with informative observation times and a dependent terminal event," *Journal of the American Statistical Association*, 107, 688–700.

Wang, M. C., Qin, J., and Chiang, C. T. (2001), "Analyzing recurrent event data with informative censoring," *Journal of the American Statistical Association,* 96, 1057–1065.

Zhao, X., Deng, S., Liu, L., and Liu, L. (2014), "Sieve estimation in semiparametric modeling of longitudinal data with informative observation times," *Biostatistics,* 15, 140–153.

Zhao, X., Tong, X., and Sun, L. (2012), "Joint analysis of longitudinal data with dependent observation times," *Statistica Sinica*, 22, 317–336.

Zhou, J., Zhao, X., and Sun, L. (2013), "A new inference approach for joint models of longitudinal data with informative observation and censoring times," *Statistica Sinica*, 23, 571–593.

Zhu, L., and Fang, K. (1996), "Asymptotic for kernel estimate of sliced inverse regression," *The Annals of Statistics*, 24, 1053–1068.

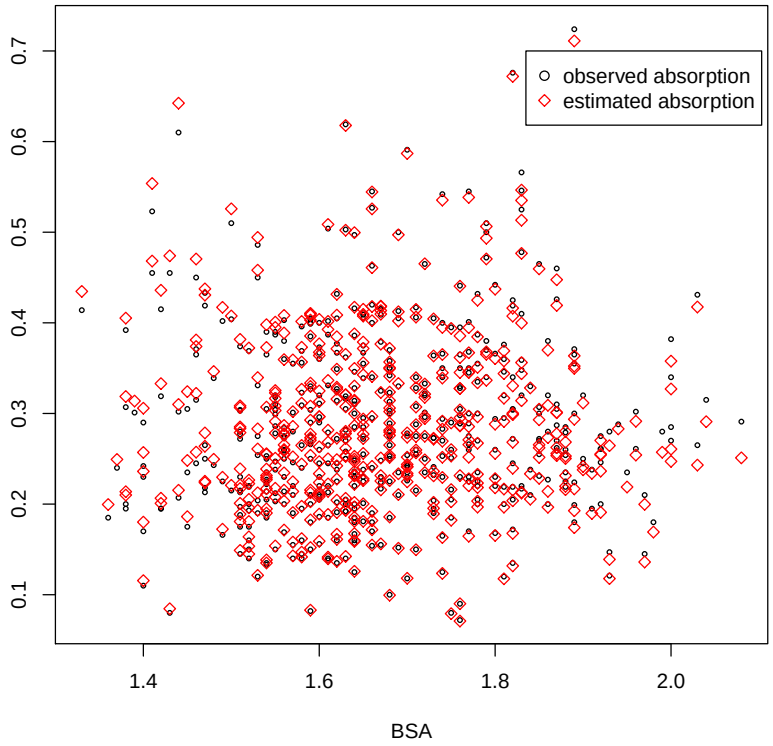

Figure 1: Estimated and observed calcium absorption.

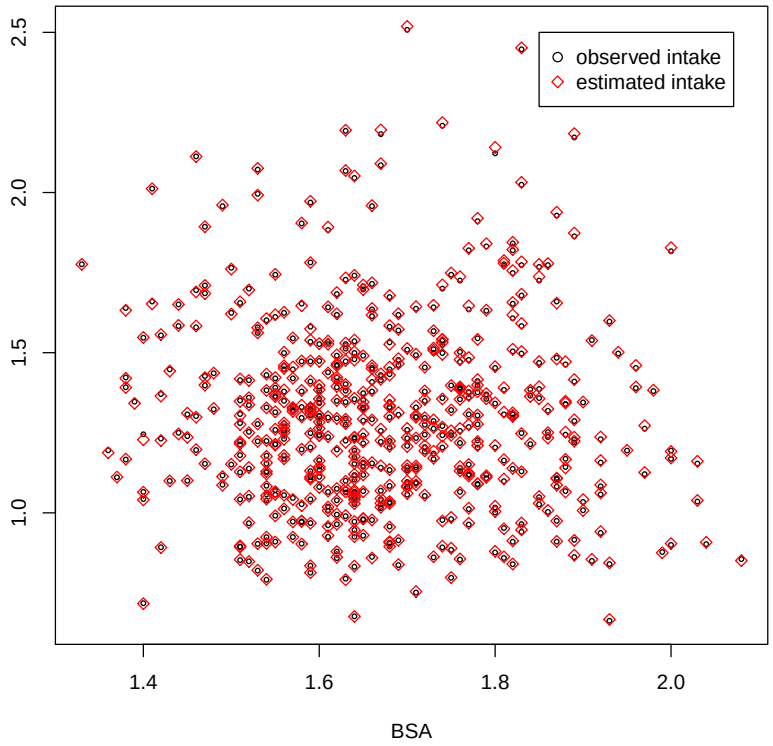

Figure 2: Estimated and observed calcium intake.

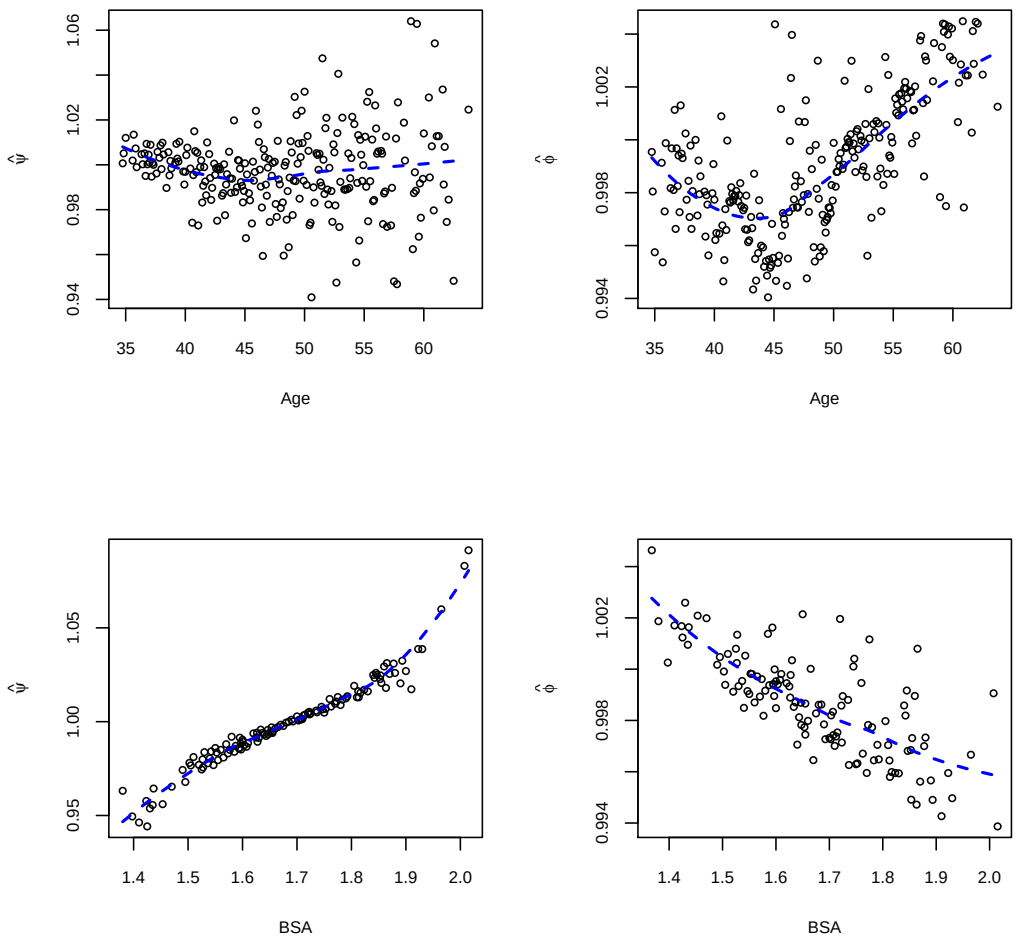

Figure 3: The top two panels are scatter plots with the loess fitted curves (blue dotted line) for $\hat{\psi}_U(\cdot)$ (top left) and $\hat{\phi}_U(\cdot)$ (top right) along with distinct observation times (ages); the bottom two panels are scatter plots with the loess fitted curves (blue dotted line) for $\hat{\psi}_T(\cdot)$ (bottom left) and $\hat{\phi}_T(\cdot)$ (bottom right) along with distinct confounding covariate (BSA) values. The plots are given after deleting the minimum and maximum points.

Table 1: Simulation results for β under the covariate-independent censoring and time-independent covariate situation.

ρ	β	$n = 100$				$n = 200$			
		BIAS	SSE	ESE	CP	BIAS	SSE	ESE	CP
-0.5	-1	0.0410	0.4109	0.4067	0.9450	0.0566	0.2702	0.2797	0.9490
		0.1455	0.4183			0.1384	0.2725		
	0	0.0400	0.4075	0.3902	0.9400	0.0514	0.2686	0.2772	0.9510
		0.1292	0.4162			0.1227	0.2705		
	1	0.0356	0.4074	0.3923	0.9380	0.0454	0.2679	0.2787	0.9530
		0.1129	0.4149			0.1069	0.2691		
0	-1	0.0338	0.3979	0.3849	0.9450	0.0570	0.2610	0.2729	0.9510
		0.1403	0.4062			0.1406	0.2635		
	0	0.0320	0.3962	0.3807	0.9430	0.0517	0.2596	0.2706	0.9500
		0.1240	0.4041			0.1249	0.2614		
	1	0.0275	0.3963	0.3828	0.9420	0.0457	0.2591	0.2720	0.9540
		0.1076	0.4029			0.1092	0.2599		
0.5	-1	0.0260	0.4092	0.3941	0.9450	0.0573	0.2682	0.2795	0.9470
		0.1351	0.4164			0.1428	0.2711		
	0	0.0240	0.4080	0.3899	0.9360	0.0520	0.2671	0.2773	0.9480
		0.1188	0.4145			0.1271	0.2690		
	1	0.0196	0.4084	0.3918	0.9360	0.0460	0.2666	0.2788	0.9490
		0.1024	0.4135			0.1114	0.2675		

Note: the first row is for the proposed estimates, and the second row is for the competing estimates disregarding the distortion fact in the underlying response and predictors.

Table 2: Simulation results for β under the covariate-independent censoring and time-dependent covariate situation.

ρ	β	$n = 100$				$n = 200$			
		BIAS	SSE	ESE	CP	BIAS	SSE	ESE	CP
-0.5	-1	0.0139	0.4067	0.4704	0.9620	-0.0221	0.3551	0.3784	0.9550
		0.0109	0.4113			-0.0048	0.3593		
	0	0.0504	0.3960	0.4517	0.9620	-0.0024	0.2844	0.3328	0.9710
		0.0890	0.3914			0.0650	0.2793		
	1	0.0851	0.4838	0.4681	0.9330	-0.0025	0.3295	0.3408	0.9500
		0.1670	0.5121			0.1402	0.3725		
0	-1	0.0125	0.4027	0.4595	0.9610	-0.0048	0.2896	0.3376	0.9650
		0.0085	0.4074			-0.0091	0.2881		
	0	0.0482	0.3938	0.4430	0.9640	-0.0003	0.2748	0.3255	0.9650
		0.0865	0.3900			0.0662	0.2709		
	1	0.0829	0.4830	0.4595	0.9290	0.0002	0.3229	0.3336	0.9470
		0.1646	0.5131			0.1415	0.3693		
0.5	-1	0.0104	0.4182	0.4651	0.9570	-0.0040	0.2905	0.3388	0.9690
		0.0061	0.4211			-0.0079	0.2902		
	0	0.0459	0.4107	0.4495	0.9610	0.0030	0.2808	0.3289	0.9660
		0.0841	0.4069			0.0674	0.2772		
	1	0.0806	0.4979	0.4656	0.9240	0.0030	0.3296	0.3370	0.9470
		0.1622	0.5280			0.1427	0.3770		

Note: the first row is for the proposed estimates, and the second row is for the competing estimates disregarding the distortion fact in the underlying response and predictors.

Table 3: Simulation results for β under the covariate-dependent censoring and time-independent covariate situation.

ρ	β	$n = 100$				$n = 200$			
		BIAS	SSE	ESE	CP	BIAS	SSE	ESE	CP
-0.5	-1	0.0314	0.3981	0.3880	0.9280	0.0306	0.2710	0.2760	0.9550
		0.1358	0.3989			0.1239	0.2737		
	0	0.0300	0.3958	0.3845	0.9260	0.0253	0.2706	0.2741	0.9550
		0.1166	0.3980			0.1051	0.2723		
	1	0.0256	0.3960	0.3871	0.9300	0.0310	0.2766	0.2770	0.9450
		0.0974	0.3981			0.0957	0.2767		
0	-1	0.0236	0.3838	0.3781	0.9400	0.0321	0.2606	0.2695	0.9520
		0.1296	0.3838			0.1276	0.2653		
	0	0.0216	0.3820	0.3748	0.9370	0.0139	0.2605	0.2679	0.9540
		0.1104	0.3829			0.1087	0.2637		
	1	0.0170	0.3823	0.3774	0.9370	0.0360	0.2724	0.2704	0.9470
		0.0913	0.3830			0.1030	0.2718		
0.5	-1	0.0152	0.3935	0.3880	0.9410	0.0520	0.2857	0.2776	0.9410
		0.1235	0.3913			0.1509	0.2860		
	0	0.0132	0.3915	0.3846	0.9400	0.0283	0.2677	0.2748	0.9550
		0.1043	0.3904			0.1123	0.2728		
	1	0.0086	0.3919	0.3970	0.9430	0.0226	0.2679	0.2766	0.9560
		0.0851	0.3906			0.0935	0.2719		

Note: the first row is for the proposed estimates, and the second row is for the competing estimates disregarding the distortion fact in the underlying response and predictors.

Table 4: Simulation results for β under the covariate-dependent censoring and time-dependent covariate situation.

ρ	β	$n = 100$				$n = 200$			
		BIAS	SSE	ESE	CP	BIAS	SSE	ESE	CP
-0.5	-1	0.0080	0.4437	0.3769	0.9660	-0.0211	0.2855	0.3198	0.9680
		0.0183	0.3785			-0.0230	0.2873		
	0	0.0316	0.3628	0.4251	0.9710	-0.0048	0.2752	0.3054	0.9600
		0.0716	0.3635			0.0451	0.2735		
	1	0.0441	0.4219	0.4354	0.9470	0.0036	0.3193	0.3107	0.9380
		0.1249	0.4559			0.1133	0.3629		
0	-1	0.0050	0.3674	0.4329	0.9670	-0.0168	0.2753	0.3089	0.9660
		0.0126	0.3700			-0.0224	0.2798		
	0	0.0271	0.3554	0.4164	0.9710	-0.0032	0.2655	0.2978	0.9620
		0.0658	0.3565			0.0458	0.2656		
	1	0.0396	0.4166	0.4269	0.9470	0.0052	0.3109	0.3032	0.9420
		0.1191	0.4517			0.1139	0.3569		
0.5	-1	0.0012	0.3789	0.4382	0.9690	-0.0155	0.2778	0.3112	0.9730
		0.0068	0.3799			-0.0217	0.2840		
	0	0.0228	0.3682	0.4223	0.9680	-0.0016	0.2688	0.3005	0.9630
		0.0601	0.3684			0.0464	0.2699		
	1	0.0352	0.4286	0.4327	0.9510	0.0068	0.3135	0.3058	0.9350
		0.1134	0.4625			0.1146	0.3601		

Note: the first row is for the proposed estimates, and the second row is for the competing estimates disregarding the distortion fact in the underlying response and predictors.