Modelling survival events with longitudinal covariates measured with error

Hongsheng Dai, University of Brighton, UK

E–mail: h.dai@brighton.ac.uk

Jianxin Pan, University of Manchester, UK

Yanchun Bao[*], Yunnan Normal University, China

Abstract

In survival analysis, time-dependent covariates are usually present as longitudinal data collected periodically and measured with error. The longitudinal data can be assumed to follow a linear mixed effect model and Cox regression models may be used for modelling of survival events. The hazard rate of survival times may depend on the underlying time-dependent covariates measured with error, which may be described by random effects. Most existing methods proposed for such models assume a parametric distribution assumption on the random effects and specify a normally distributed error term for the linear mixed effect model. These assumptions may not be always valid in practice. In this paper we propose a new likelihood method for Cox regression models with error-contaminated time-dependent covariates. The proposed method does not require

[*] Supported by grant Z2006-1-65010 from Chun Hui Plan Scientific Research Project of the Ministry of Education of the People's Republic of China.

any parametric distribution assumptions on the random effects and random errors. Asymptotic properties for parameter estimators are provided. Simulation results show that the proposed method is more efficient than the existing methods.

Key words: Censored data; Longitudinal measurements; Partial likelihood; Proportional hazard model.

1 Introduction

Cox proportional hazard model (Cox, 1972) is widely used to study the relationship between survival events and time-dependent or time-independent covariates, which assumes that the failure time hazard rate function $\lambda(s)$ relates to covariates through

$$\lambda(s) \quad = \quad \lambda_0(s)\exp(\gamma W(s)).$$

To implement the above Cox model most existing methods require the time dependent covariate process $W(s)$ to be fully observed. In practice, however, $W(s)$ is measured intermittently and very likely with error. In other words, we only observe longitudinal measurements $\widetilde{W}_j$ at some time points $t_j, j = 1, \cdots, m$, where $\widetilde{W}_j = W(t_j) + \epsilon_j$ and ϵ_j is the error term. Substituting mis-measured values for true covariates in Cox models can lead to very biased estimates (Prentice, 1982).

Recent studies focus on joint modelling of survival events and longitudinal measurements. The latent time-dependent process $W(s)$ is usually assumed to be $W(s) = \omega_0 + \omega_1 s$, which may be generalized to a general polynomial in time. Here ω_0 and ω_1 are random effects. Such assumptions can be used to study the effects of potentially mis-measured time-dependent covariates on the failure time. Many existing methods (DeGruttola and Tu, 1994; Faucett and Thomas, 1996; Henderson et.al., 2000; Wulfson and Tsiatis, 1997) assumed that the random effects and the random error follow Gaussian distributions, and EM algorithms or Bayesian approaches were used to deal with the unobserved covariate process $W(s)$. However, the normal distribution assumption on random effects and random errors may not be always true in practice. Misspecification of the distributions of random effects or random errors can lead to very biased estimates (Tisiatis and Davidian, 2001; Song and Huang, 2005; Wang, 2006).

Hu et.al. (1998) and Song et.al. (2002b) relaxed the normal assumption by assuming that the density of the underlying covariates belongs to a smooth class. These approaches involve an intensive computation of the use of EM-algorithm. Huang and Wang (2000) and Song and Huang (2005) proposed a corrected score approach which does not require any distribution assumption on the random effects and the error term ϵ. Their models assume that the underlying covariate W is time-independent. In addition, the corrected score method requires replicated covariate observations for each subject. In many applications, however, the underlying covariates are time-dependent and replicated covariate observations for each subject may not be available. Recently, an interesting method was proposed by Wang (2006). With the normal distribution assumption for the error terms, Wang' method does not have any distribution assumption on $W(s)$. This method is based on the assumption that the hazard rate depends on the time-independent random effects, not the time-dependent underlying process $W(s)$. Another estimation method, called conditional score (CS) estimator, was proposed by Tisiatis and Davidian (2001), where the hazard rate $\lambda(s)$ is assumed to depend on the time-dependent underlying process and the error term is assumed to be normally distributed. Note that the CS estimator does not require any distribution assumption on $W(s)$.

In this paper, following Tisiatis and Davidian (2001) we assume that the failure time hazard rate depends on a time-dependent covariate process. With the normal distribution assumption on the error ϵ, a simple working-likelihood (SWL) estimator is proposed without involving any distribution assumption on ω_0 and ω_1. The SWL estimator is proved to be consistent and asymptotically normally distributed under some regular conditions. A consistent covariance estimator is also provided. We then relax the normal distribution assumption on the error ϵ. A generalized working-likelihood (GWL) estimator is proposed for such cases. Consistency and asymptotic normality of the GWL estimator are provided. Simulation studies demonstrate that when the error term ϵ follows a normal distribution, the SWL estimator is as efficient as the CS estimator of Tisiatis and Davidian (2001). Numerical studies also show that the GWL estimator works very well and it is more efficient than the SWL and CS estimators when either ϵ or (ω_0, ω_1) is not normally distributed. This paper is organized as follows. Models and notation are given in Section 2. In Section 3 we propose the simple working-likelihood estimator. The generalized working-likelihood estimator is introduced in Section 4. Numerical studies, real data analysis and discussions are given in Sections 5 and 6.

2 Models and notations

Let $\boldsymbol{Z}_i(s)$ be the observed time-dependent covariate and $W_i(s)$ be the unobserved time-dependent process. Throughout this paper for simplicity we assume that $W_i(s)$ is a univariate process, though the proposed methods can be extended to a multivariate unobserved time-dependent process. Without loss of generality, we assume that $W_i(s) = \sum_{j=0}^{q-1} \omega_{ij} s^j$. Here $\boldsymbol{\omega}_i = (\omega_{i0}, \cdots, \omega_{i,q-1})$ is the random coefficients/effects for the ith subject. In practice we cannot observe $W_i(s)$ but observe m_i longitudinal measurements $\widetilde{\boldsymbol{W}}_i = \{\widetilde{W}_{ij}, j = 1, \cdots, m_i\}$ as

$$\widetilde{W}_{ij} \quad = \quad W_i(t_{ij}) + \epsilon_{ij}, \tag{1}$$

at ordered times $\boldsymbol{t}_i = (t_{i1}, \cdots, t_{i,m_i})^T$. We assume that $\boldsymbol{\omega}_i$ is independent of $\boldsymbol{t}_i$. Let $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \cdots, \epsilon_{i,m_i})$. Throughout this paper we do not put any distribution assumption on $\boldsymbol{\omega}_i$.

Let T_i be the survival time of the ith subject. In practice, we may not observer T_i for all subjects. Instead, we only observe $\tilde{T}_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \le C_i)$, where the censoring variable C_i is independent of T_i. There is an additional censoring at τ, which is the end time of the experiment. We assume that T_i and C_i are independent of $\boldsymbol{t}_i$ and $\boldsymbol{\epsilon}_i$. Cox proportional hazard model assumes that the hazard rate is a function of covariates through the following form, $\lambda_i(s) = \lambda_0(s) \exp(\gamma W_i(s) + \boldsymbol{\beta}^T \boldsymbol{Z}_i(s))$ where $\lambda_0(t)$ is an arbitrary baseline hazard function.

Define counting processes $dN_i(s) = I[s \le \tilde{T}_i \le s + ds, \delta_i = 1, t_{i,m_i} \le s]$ and at-risk processes $Y_i(s) = I[\tilde{T}_i \ge s, t_{i,m_i} \le s]$. We have

$$\mathcal{E}(dN_i(s)|\mathcal{F}_{i,s}) \quad = \quad \lambda_0(s) ds \exp(\gamma W_i(s) + \boldsymbol{\beta}^T \boldsymbol{Z}_i(s)) Y_i(s), \tag{2}$$

where $\mathcal{F}_{i,s}$ is the filtration generated from σ-fields $\sigma\{\tilde{T}_i \le u, \delta_i, \boldsymbol{\omega}_i, \boldsymbol{Z}_i(u), u \le s\}$ and

$\sigma\{t_{i,m_i} \leq u, u \leq \tau\}$. The log-partial likelihood function for parameter $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta}^T)^T$ is

$$l(\boldsymbol{\theta}) \quad = \quad n^{-1} \sum_i \int \left[\gamma W_i(s) + \boldsymbol{\beta}^T \boldsymbol{Z}_i(s) - \log E^{(0)}(W, \boldsymbol{\theta}, s)\right] dN_i(s), \tag{3}$$

where $E^{(0)}(W, \boldsymbol{\theta}, s) = n^{-1} \sum_i E_i^{(0)}(W_i, \boldsymbol{\theta}, s) := n^{-1} \sum_i \exp[\gamma W_i(s) + \boldsymbol{\beta}^T \boldsymbol{Z}_i(s)] Y_i(s)$. If $W_i(s)$ is fully observed, then we can maximize $l(\boldsymbol{\theta})$ by solving the following equations

$$\boldsymbol{U}^{(1)}(\boldsymbol{\theta}, \tau) \quad := \quad n^{-1} \sum_i \int_0^{\tau} \left[\begin{pmatrix} W_i(s) \\ \boldsymbol{Z}_i(s) \end{pmatrix} - \frac{\boldsymbol{E}^{(1)}(W, \boldsymbol{\theta}, s)}{E^{(0)}(W, \boldsymbol{\theta}, s)} \right] dN_i(s) = \mathbf{0}, \tag{4}$$

where $\boldsymbol{E}^{(1)}(W, \boldsymbol{\theta}, s) = \frac{\partial E^{(0)}(W, \boldsymbol{\theta}, s)}{\partial (\gamma, \boldsymbol{\beta}^T)^T}$ and

$$\boldsymbol{E}^{(1)}(W, \boldsymbol{\theta}, s) = n^{-1} \sum_i \boldsymbol{E}_i^{(1)}(W_i, \boldsymbol{\theta}, s) := n^{-1} \sum_i \left(W_i(s), \boldsymbol{Z}_i(s)^T\right)^T E_i^{(0)}(W_i, \boldsymbol{\theta}, s).$$

Using martingale theories, under some regular conditions it is straightforward to show that the maximum likelihood estimate based on (3) is consistent.

When $W_i(s)$ is measured with error, the score function in (4) is not available to use. To solve this problem, a naive approach is to replace $W_i(s)$ with its least square estimate and then treat this estimate as a covariate. Another method is to use the regression calibration estimate, which replaces $W_i(s)$ with its conditional expectation given the longitudinal measurements. These approaches, however, result in a severe bias to the estimate of γ. Detailed discussions and comparisons can be found in Tisiatis and Davidian (2001) and Wang (2006). Based on a sufficient statistic for $W_i(s)$ Tisiatis and Davidian (2001) proposed a conditional score estimator, where no distribution about the random effects $\boldsymbol{\omega}_i$ is assumed. The conditional score estimate is more efficient than the naive approach and regression calibration.

3 A simple working likelihood estimator

Throughout this section, we assume that $\boldsymbol{\epsilon}_i$ in (1) has a normal distribution $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_{m_i})$. We assume that $\boldsymbol{\omega}_i, i = 1, \cdots, n$ are i.i.d. with an unknown common multivariate distribution.

3.1 Unbiased working log-likelihood function

Let $\widehat{W}_i(s)$ be the ordinary LSE of $W_i(s)$ using all the longitudinal observations $\widetilde{\boldsymbol{W}}_i$. This requires at least q longitudinal measurements on subject i. Denote $\boldsymbol{s} = (1, s, \cdots, s^{q-1})^T$ and

$$\boldsymbol{A}_i = \begin{pmatrix} 1 & t_{i,1} & \cdots & t_{i,1}^{q-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{i,m_i} & \cdots & t_{i,m_i}^{q-1} \end{pmatrix}.$$

Then we have $\text{Var}(\widehat{W}_i(s)|W_i(s)) = \sigma^2 v_i(s)$ where $v_i(s) := \boldsymbol{s}^T(\boldsymbol{A}_i^T \boldsymbol{A}_i)^{-1}\boldsymbol{s}$. A consistent estimator (Tisiatis and Davidian, 2001) for σ^2 is $\hat{\sigma}^2 = \frac{\sum_i I[m_i>q]R_i}{\sum_i I[m_i>q](m_i-q)}$, where R_i is the residual sum of squares for subject i based on the least squares fit to all the m_i observations.

Let $\hat{E}^{(0)}(\boldsymbol{\theta}, \sigma^2, s) = n^{-1}\sum_i \hat{E}_i^{(0)}(\boldsymbol{\theta}, \sigma^2, s)$ and

$$\hat{E}_i^{(0)}(\boldsymbol{\theta}, \sigma^2, s) := \exp\left[\gamma\widehat{W}_i(s) - \frac{\gamma^2\sigma^2 v_i(s)}{2} + \boldsymbol{\beta}^T \boldsymbol{Z}_i(s)\right] Y_i(s).$$

It is obvious that given $W_i(s)$ the LSE $\widehat{W}_i(s)$ is normally distributed. Thus given $W_i(s)$ the random variable $\exp[\gamma\widehat{W}_i(s)]$ has a log-normal distribution. Using the well-known results for the expectation of a log-normal distribution, we have $\mathcal{E}\{\exp[\gamma\widehat{W}_i(s)]|W_i(s)\} = \exp[\gamma W_i + \gamma^2\sigma^2 v_i(s)/2]$, where $\mathcal{E}$ represents the expectation. Therefore

$$\mathcal{E}[\hat{E}_i^{(0)}(\boldsymbol{\theta}, \sigma^2, s)|W_i(s)] \quad = \quad E_i^{(0)}(W_i, \boldsymbol{\theta}, s) \tag{5}$$

which implies $\mathcal{E}\hat{E}_i^{(0)}(\boldsymbol{\theta}, \sigma^2, s) = \mathcal{E}E_i^{(0)}(W_i, \boldsymbol{\theta}, s)$.

Under some regular conditions (see Appendix A, **C.2**), according to (5) we can show that, in probability,

$$\lim_{n\to\infty} \hat{E}^{(0)}(\boldsymbol{\theta}, \sigma^2, s) = \lim_{n\to\infty} E^{(0)}(W, \boldsymbol{\theta}, s) := e^{(0)}(\boldsymbol{\theta}, s). \tag{6}$$

Assume σ^2 is known first. We consider the following working likelihood function

$$\hat{l}_n(\boldsymbol{\theta}, \sigma^2) = n^{-1}\sum_i \int \left[\gamma\widehat{W}_i(s) + \boldsymbol{\beta}^T \boldsymbol{Z}_i(s) - \log \hat{E}^{(0)}(\boldsymbol{\theta}, \sigma^2, s)\right] dN_i(s). \tag{7}$$

From (6) we know that $\hat{l}_n(\boldsymbol{\theta}, \sigma^2)$ is a working likelihood function asymptotically unbiased to $l(\boldsymbol{\theta})$ given in (3).

3.2 The maximum likelihood estimator

To maximize $\hat{l}_n(\boldsymbol{\theta}, \sigma^2)$ for the fixed σ^2, given in (7), we solve the following equations,

$$\begin{aligned} \hat{U}_\gamma^{(1)}(\boldsymbol{\theta}, \sigma^2, \tau) \quad &:= \quad n^{-1}\sum_i \int_0^\tau \left[\widehat{W}_i(s) - \frac{\hat{E}_\gamma^{(1)}(\boldsymbol{\theta}, \sigma^2, s)}{\hat{E}^{(0)}(\boldsymbol{\theta}, \sigma^2, s)}\right] dN_i(s) = 0, \\ \hat{\boldsymbol{U}}_{\boldsymbol{\beta}}^{(1)}(\boldsymbol{\theta}, \sigma^2, \tau) \quad &:= \quad n^{-1}\sum_i \int_0^\tau \left[\boldsymbol{Z}_i(s) - \frac{\hat{\boldsymbol{E}}_{\boldsymbol{\beta}}^{(1)}(\boldsymbol{\theta}, \sigma^2, s)}{\hat{E}^{(0)}(\boldsymbol{\theta}, \sigma^2, s)}\right] dN_i(s) = \boldsymbol{0}, \end{aligned} \tag{8}$$

where

$$\hat{E}^{(1)}_{\gamma}(\boldsymbol{\theta}, \sigma^2, s) = \frac{\partial \hat{E}^{(0)}(\boldsymbol{\theta}, \sigma^2, s)}{\partial \gamma} = n^{-1} \sum_i \left[\widehat{W}_i(s) - \sigma^2 v_i(s)\gamma\right] \hat{E}^{(0)}_i(\boldsymbol{\theta}, \sigma^2, s),$$
$$\hat{\boldsymbol{E}}^{(1)}_{\boldsymbol{\beta}}(\boldsymbol{\theta}, \sigma^2, s) = \frac{\partial \hat{E}^{(0)}(\boldsymbol{\theta}, \sigma^2, s)}{\partial \boldsymbol{\beta}} = n^{-1} \sum_i \boldsymbol{Z}_i(s) \hat{E}^{(0)}_i(\boldsymbol{\theta}, \sigma^2, s).$$

Let $\hat{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}, \sigma^2, s) = (\hat{E}^{(1)}_{\gamma}(\boldsymbol{\theta}, \sigma^2, s), \hat{\boldsymbol{E}}^{(1)}_{\boldsymbol{\beta}}(\boldsymbol{\theta}, \sigma^2, s)^T)^T$ and

$$\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}, \sigma^2, s) = (\hat{U}^{(1)}_{\gamma}(\boldsymbol{\theta}, \sigma^2, s), \hat{\boldsymbol{U}}^{(1)}_{\boldsymbol{\beta}}(\boldsymbol{\theta}, \sigma^2, s)^T)^T.$$

Then we can write the estimating equations in (8) as

$$\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}, \sigma^2, \tau) \ := \ n^{-1} \sum_i \int_0^{\tau} \left[\begin{pmatrix} \widehat{W}_i(s) \\ \boldsymbol{Z}_i(s) \end{pmatrix} - \frac{\hat{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}, \sigma^2, s)}{\hat{E}^{(0)}(\boldsymbol{\theta}, \sigma^2, s)} \right] dN_i(s) = \boldsymbol{0}. \qquad (9)$$

If σ^2 is unknown, it can be replaced by the consistent estimator $\hat{\sigma}^2$. In this case, $\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}, \hat{\sigma}^2, \tau) = 0$ is an estimating equation asymptotically unbiased to (4).

Theorem 3.1. *Let $\hat{\boldsymbol{\theta}}$ be the estimated value by solving $\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}, \hat{\sigma}^2, \tau) = \boldsymbol{0}$ and $\boldsymbol{\theta}_0$ be the true parameter. Under some regular conditions given in Appendix A, we have* $\lim_{n\to\infty} \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$ *in probability.*

Proof. Under some regular conditions $l(\boldsymbol{\theta})$ is concave. The theorem then follows from the facts that $\hat{l}_n(\boldsymbol{\theta}, \sigma^2)$ is asymptotically unbiased to $l(\boldsymbol{\theta})$ and that under some mild regular conditions $l(\boldsymbol{\theta})$ has a unique maximum at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. □

Let $\bar{N}(s) = \sum_i N_i(s)/n$. A consistent estimator for $\lambda_0(s)ds$ is $\hat{\lambda}_0(s)ds = \frac{d\bar{N}(s)}{\hat{E}^{(0)}(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2, s)}$.

3.3 Covariance estimation

Let

$$
\begin{aligned}
\hat{\boldsymbol{E}}^{(2)}(\boldsymbol{\theta},\sigma^2,s) &:= \frac{\partial \hat{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta},\sigma^2,s)}{\partial \boldsymbol{\theta}} \\
&= n^{-1}\sum_i \begin{pmatrix} [\widehat{W}_i(s)-\sigma^2 v_i(s)\gamma]^2-\sigma^2 v_i(s) & [\widehat{W}_i(s)-\sigma^2 v_i(s)\gamma]\boldsymbol{Z}_i(s)^T \\ \boldsymbol{Z}_i(s)[\widehat{W}_i(s)-\sigma^2 v_i(s)\gamma] & \boldsymbol{Z}_i(s)\boldsymbol{Z}_i(s)^T \end{pmatrix} \hat{E}_i^{(0)}(\boldsymbol{\theta},\sigma^2,s)
\end{aligned}
$$

and

$$
\hat{\boldsymbol{V}}(\boldsymbol{\theta},\sigma^2,s) := \frac{\hat{\boldsymbol{E}}^{(2)}(\boldsymbol{\theta},\sigma^2,s)}{\hat{E}^{(0)}(\boldsymbol{\theta},\sigma^2,s)} - \left\{\frac{\hat{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta},\sigma^2,s)}{\hat{E}^{(0)}(\boldsymbol{\theta},\sigma^2,s)}\right\}^{\otimes 2}
$$

where for any vector $\boldsymbol{a}$ the notation $\boldsymbol{a}^{\otimes 2}$ represents the outer product $\boldsymbol{a}\boldsymbol{a}'$.

Under regular conditions **C.2** and **C.3** in Appendix A, $\hat{\boldsymbol{V}}(\boldsymbol{\theta},\sigma^2,s)$ converges uniformly to $\boldsymbol{v}(\boldsymbol{\theta},s)$ given in **C.3** in Appendix A and $n^{-1}\int \hat{\boldsymbol{V}}(\boldsymbol{\theta},\hat{\sigma}^2,s)\sum_i dN_i(s)$ converges to $\int \boldsymbol{v}(\boldsymbol{\theta},s)e^{(0)}(\boldsymbol{\theta},s)\lambda_0(s)ds$ in probability. The asymptotic normality of $\hat{\boldsymbol{\theta}}$ is established by the following theorem.

Theorem 3.2. *For the estimator $\hat{\boldsymbol{\theta}}$ in Theorem 1, we have $n^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \Rightarrow N(\boldsymbol{0},\boldsymbol{R})$ where $\boldsymbol{R}=\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0,\tau)^{-1}\boldsymbol{\Sigma}_U(\boldsymbol{\theta}_0,\sigma^2,\tau)\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0,\tau)^{-1}$, $\Sigma_U(\boldsymbol{\theta}_0,\sigma^2,\tau)=\lim_{n\to\infty} Var[\sqrt{n}\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta},\hat{\sigma}^2,\tau)]$ and*

$$
\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0,\tau) = -\int \boldsymbol{v}(\boldsymbol{\theta}_0,s)e^{(0)}(\boldsymbol{\theta}_0,s)\lambda_0(s)ds.
$$

Note that a consistent estimator for $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0,\tau)$ is

$$
\hat{\boldsymbol{\mathcal{I}}}(\hat{\boldsymbol{\theta}},\hat{\sigma}^2,\tau) = -\frac{1}{n}\int\sum_i \hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}},\hat{\sigma}^2,s)dN_i(s).
$$

Define

$$\begin{aligned}\hat{\boldsymbol{\Phi}}_i(\boldsymbol{\theta}_0,\hat{\sigma}^2,t) \;&=\; \int_0^t \left[\begin{pmatrix}\widehat{W}_i(s)\\ \boldsymbol{Z}_i(s)\end{pmatrix} - \frac{\hat{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}_0,\hat{\sigma}^2,s)}{\hat{E}^{(0)}(\boldsymbol{\theta}_0,\hat{\sigma}^2,s)}\right] dN_i(s) \qquad (10)\\ &\quad -\int_0^t \left[\begin{pmatrix}\widehat{W}_i(s)-\hat{\sigma}^2 v_i(s)\gamma\\ \boldsymbol{Z}_i(s)\end{pmatrix} - \frac{\hat{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}_0,\hat{\sigma}^2,s)}{\hat{E}^{(0)}(\boldsymbol{\theta}_0,\hat{\sigma}^2,s)}\right] \hat{E}_i^{(0)}(\boldsymbol{\theta}_0,\hat{\sigma}^2,s)\hat{\lambda}_0(s)ds.\end{aligned}$$

A consistent estimator for $\boldsymbol{\Sigma}_U(\boldsymbol{\theta}_0,\sigma^2,t)$ is

$$\hat{\boldsymbol{\Sigma}}_U(\hat{\boldsymbol{\theta}},\hat{\sigma}^2,t) = \frac{1}{n}\sum_{i}^{n}\hat{\boldsymbol{\Phi}}_i(\hat{\boldsymbol{\theta}},\hat{\sigma}^2,t)^{\otimes 2} + \sum_{i=2}^{n}\hat{\boldsymbol{\Phi}}_i(\hat{\boldsymbol{\theta}},\hat{\sigma}^2,t)\otimes\hat{\boldsymbol{\Phi}}_1(\hat{\boldsymbol{\theta}},\hat{\sigma}^2,t). \qquad (11)$$

Thus a consistent estimator for $\boldsymbol{R}$ is $\hat{\boldsymbol{R}} = \hat{\boldsymbol{\mathcal{I}}}(\hat{\boldsymbol{\theta}},\hat{\sigma}^2,\tau)^{-1}\hat{\boldsymbol{\Sigma}}_U(\hat{\boldsymbol{\theta}},\hat{\sigma}^2,\tau)\hat{\boldsymbol{\mathcal{I}}}(\hat{\boldsymbol{\theta}},\hat{\sigma}^2,\tau)^{-1}$.

4 A general working likelihood estimator

Throughout this section, we relax the normal distribution assumption for the random errors $\boldsymbol{\epsilon}_i$ and only assume $\epsilon_{ij}, j=1,\cdots,m_i, i=1,\cdots,n$ are i.i.d. with mean 0 and finite second moments. We also assume that given $\boldsymbol{\omega}_i$, $(\widetilde{W}_{ij},t_{ij}), j=1,\cdots,m_i$ are i.i.d. pairs.

4.1 Unbiased log-likelihood function and unbiased estimating equation

Let $\xi_i(s)=\widehat{W}_i(s)-W_i(s)$. We then have

$$\xi_i(s)=\boldsymbol{s}^T(\boldsymbol{A}_i^T\boldsymbol{A}_i)^{-1}\boldsymbol{A}_i^T\boldsymbol{\epsilon}_i. \qquad (12)$$

Note that $\xi_i(s), i = 1, \cdots, n$ may not have the same distribution since the number of longitudinal measurements of each subject, m_i, may not be necessarily the same.

Suppose each subject has M extra longitudinal observations, i.e., $(\widetilde{W}_{ij,1}, t_{ij,1})$ $j = 1, \cdots, M$, which are i.i.d pairs and are also independent of $(\widetilde{W}_{ij}, t_{ij}), j = 1, \cdots, m_i$. The Least-Squares estimator based on the extra longitudinal observations is denoted by $\widehat{W}_{i,1}(s)$. Let $\xi_{i,1}(s) = \widehat{W}_{i,1}(s) - W_i(s)$. Since $\xi_{i,1}(s)$ has a similar expression as that in (12) and each subject has M replicated longitudinal measurements, we know that $\xi_{i,1}(s), i = 1, \cdots, n$ are i.i.d. random variables.

Let $\varphi^{(k)}(\gamma, s) = \mathcal{E}[\xi_{i,1}(s)^k \exp(\gamma \xi_{i,1}(s))]$ for $k = 0, 1, 2$. Denote

$$\breve{E}_i^{(0)}(\boldsymbol{\theta}, s) = E_i^{(0)}(\widehat{W}_{i,1}, \boldsymbol{\theta}, s).$$

Note that $\breve{E}_i^{(0)}(\boldsymbol{\theta}, s)$ is $E_i^{(0)}(W_{i,1}, \boldsymbol{\theta}, s)$ with W_i replaced by $\hat{W}_{i,1}$, the Least Squares estimator based on the M extra longitudinal observations.

Since $\xi_i(s) - \xi_{i,1}(s) = \widehat{W}_i(s) - \widehat{W}_{i,1}(s)$, we have

$$\begin{aligned}
\frac{\breve{E}_i^{(0)}(\boldsymbol{\theta}, s)}{\varphi^{(0)}(\gamma, s)} &= \frac{E_i^{(0)}(\widehat{W}_i, \boldsymbol{\theta}, s)}{\exp(\gamma \xi_i(s) - \gamma \xi_{i,1}(s))} \frac{1}{\varphi^{(0)}(\gamma, s)} \\
&= \frac{E_i^{(0)}(\widehat{W}_i, \boldsymbol{\theta}, s)}{\exp(\gamma \xi_i(s))} \frac{\exp[\gamma \xi_{i,1}(s)]}{\varphi^{(0)}(\gamma, s)} = E_i^{(0)}(W_i, \boldsymbol{\theta}, s) \frac{\exp[\gamma \xi_{i,1}(s)]}{\varphi^{(0)}(\gamma, s)}.
\end{aligned}$$

Therefore $\mathcal{E}[\breve{E}_i^{(0)}(\boldsymbol{\theta}, s)/\varphi^{(0)}(\gamma, s) | W_i(s)] = E_i^{(0)}(W_i, \boldsymbol{\theta}, s)$. If let $\breve{E}^{(0)}(\boldsymbol{\theta}, s) = n^{-1} \sum_i \breve{E}_i^{(0)}(\boldsymbol{\theta}, s)$, then we have $\lim_{n \to \infty} \breve{E}^{(0)}(\boldsymbol{\theta}, s)/\varphi^{(0)}(\gamma, s) = \lim_{n \to \infty} E^{(0)}(\boldsymbol{\theta}, s) = e^{(0)}(\boldsymbol{\theta}, s)$. Similar to the results in Section 3, we have an unbiased log-likelihood function as follows

$$\breve{l}_n(\boldsymbol{\theta}) = n^{-1} \sum_i \int \left[\gamma \widehat{W}_i(s) + \boldsymbol{\beta}^T \boldsymbol{Z}_i(s) - \log \frac{\breve{E}^{(0)}(\boldsymbol{\theta}, s)}{\varphi^{(0)}(\gamma, s)} \right] dN_i(s). \tag{13}$$

Let

$$\mathbf{\Psi} = \left\{ \psi_1(\gamma, s) = \frac{\varphi^{(1)}(\gamma, s)}{\varphi^{(0)}(\gamma, s)}, \psi_2(\gamma, s) = \frac{\varphi^{(2)}(\gamma, s)}{\varphi^{(0)}(\gamma, s)} \right\}.$$

Then unbiased estimating equations are

$$\breve{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}, \mathbf{\Psi}, \tau) \quad := \quad n^{-1} \sum_i \int_0^\tau \left[\begin{pmatrix} \widehat{W}_i(s) \\ \boldsymbol{Z}_i(s) \end{pmatrix} - \frac{\breve{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}, \mathbf{\Psi}, s)}{\breve{E}^{(0)}(\boldsymbol{\theta}, s)} \right] dN_i(s) = \mathbf{0}, \qquad (14)$$

where

$$\breve{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}, \mathbf{\Psi}, s) = \frac{\partial \breve{E}^{(0)}(\boldsymbol{\theta}, s)}{\partial \boldsymbol{\theta}} \quad := \quad n^{-1} \sum_i \begin{pmatrix} \widehat{W}_{i,1}(s) - \psi_1(\gamma, s) \\ \boldsymbol{Z}_i(s) \end{pmatrix} \breve{E}_i^{(0)}(\boldsymbol{\theta}, s).$$

Let $\breve{\boldsymbol{E}}^{(2)}(\boldsymbol{\theta}, \mathbf{\Psi}, s) = \frac{\partial \breve{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}, \mathbf{\Psi}, s)}{\partial \boldsymbol{\theta}}$. We have

$$\breve{\boldsymbol{E}}^{(2)}(\boldsymbol{\theta}, \mathbf{\Psi}, s) \quad = \quad n^{-1} \sum_i \left[\begin{pmatrix} \widehat{W}_{i,1}(s) - \psi_1(\gamma, s) \\ \\ \boldsymbol{Z}_i(s) \end{pmatrix}^{\otimes 2} - \begin{pmatrix} \psi_2(\gamma, s) - \psi_1(\gamma, s)^2 & \mathbf{0} \\ \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \right] \breve{E}_i^{(0)}(\boldsymbol{\theta}, s).$$

Thus the derivative of $\breve{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}, \mathbf{\Psi}, \tau)$ with respect to $\boldsymbol{\theta}$ is

$$\breve{\boldsymbol{\mathcal{I}}}(\boldsymbol{\theta}, \mathbf{\Psi}, \tau) \quad = \quad -\frac{1}{n} \sum_i \int \breve{\boldsymbol{V}}(\boldsymbol{\theta}, \mathbf{\Psi}, s) dN_i(s)$$

where

$$\breve{\boldsymbol{V}}(\boldsymbol{\theta}, \mathbf{\Psi}, s) = \frac{\breve{\boldsymbol{E}}^{(2)}(\boldsymbol{\theta}, \mathbf{\Psi}, s)}{\breve{E}^{(0)}(\boldsymbol{\theta}, s)} - \left\{ \frac{\breve{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}, \mathbf{\Psi}, s)}{\breve{E}^{(0)}(\boldsymbol{\theta}, s)} \right\}^{\otimes 2}.$$

Note that if we replace $\mathbf{\Psi} = \{\psi_1(\gamma, s), \psi_2(\gamma, s)\}$ in (14) by its consistent estimator, then we can calculate the MLE by solving score functions $\breve{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}, \hat{\mathbf{\Psi}}, \tau) = \mathbf{0}$.

4.2 Consistent estimator for $\mathbf{\Psi}$

Suppose that for each $(\widetilde{W}_{ij,1}, t_{ij,1}), j = 1, \cdots, M$, there are other two replicated data sets $(\widetilde{W}_{ij,r}, t_{ij,r}), j = 1, \cdots, M, r = 2, 3$. Based on the two replicated data sets we can find the LSEs $\widehat{W}_{i,r}(s), r = 2, 3$. Let $\xi_{i,r}(s) = \widehat{W}_{i,r}(s) - W_i(s)$. For $r = 2, 3$, we can calculate i.i.d. values $\xi_{i,1}(s) - \xi_{i,r}(s) = \widehat{W}_i(s) - \widehat{W}_{i,r}(s), i = 1, \cdots, n$. We then have the following theorem.

Theorem 4.1. *Let*

$$\begin{aligned}
\hat{\psi}_1(\gamma, s) &:= \frac{\sum_i (\xi_{i,1}(s) - \xi_{i,2}(s)) \exp(\gamma\xi_{i,1}(s) - \gamma\xi_{i,3}(s))}{\sum_i \exp(\gamma\xi_{i,1}(s) - \gamma\xi_{i,3}(s))}, \\
\hat{\psi}_2(\gamma, s) &:= \frac{\sum_i (\xi_{i,1}(s) - \xi_{i,2}(s))^2 \exp(\gamma\xi_{i,1}(s) - \gamma\xi_{i,3}(s))}{\sum_i \exp(\gamma\xi_{i,1}(s) - \gamma\xi_{i,3}(s))} \\
&\quad - \frac{\sum_i \sigma^2 \boldsymbol{s}^T [\boldsymbol{A}_{i,2}^T \boldsymbol{A}_{i,2}]^{-1} \boldsymbol{s} \exp(\gamma\xi_{i,1}(s) - \gamma\xi_{i,3}(s))}{\sum_i \exp(\gamma\xi_{i,1}(s) - \gamma\xi_{i,3}(s))},
\end{aligned}$$

where $\boldsymbol{A}_{i,2}$ is the regressor matrix for the second replicated data set $(\widetilde{W}_{ij,2}, t_{ij,2}), j = 1, \cdots, M$ of subject i. We then have $\hat{\psi}_1(\gamma, s) \to \psi_1(\gamma, s)$ and $\hat{\psi}_2(\gamma, s) \to \psi_2(\gamma, s)$ in probability as $n \to \infty$.

With the definition of $\hat{\mathbf{\Psi}} = (\hat{\psi}_1(\gamma, s), \hat{\psi}_2(\gamma, s))$ given in the above theorem, we have the unbiased estimating equations as follows

$$\breve{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}, \hat{\mathbf{\Psi}}, \tau) \quad := \quad n^{-1} \sum_i \int_0^\tau \left[\begin{pmatrix} \widehat{W}_i(s) \\ \boldsymbol{Z}_i(s) \end{pmatrix} - \frac{\breve{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}, \hat{\mathbf{\Psi}}, s)}{\breve{E}^{(0)}(\boldsymbol{\theta}, s)} \right] dN_i(s) = \mathbf{0}. \qquad (15)$$

Similar to the proof for Theorem 3.1 we can show that the estimated value $\breve{\boldsymbol{\theta}}$ by solving (15) is consistent.

4.3 Covariance estimation

Similar to the results in Section 3, we can show that $\sqrt{n}(\breve{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges weakly to a normal distribution $N(0, \boldsymbol{R})$. Let $\breve{\lambda}_0(s)ds = d\bar{N}(s)/\breve{E}^{(0)}(\breve{\boldsymbol{\theta}}, s)$ and

$$\begin{aligned}\breve{\boldsymbol{\Phi}}_i(\boldsymbol{\theta}, \hat{\boldsymbol{\Psi}}, t) &= \int_0^t \left[\begin{pmatrix} \widehat{W}_i(s) \\ \boldsymbol{Z}_i(s) \end{pmatrix} - \frac{\breve{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}, \hat{\boldsymbol{\Psi}}, s)}{\breve{E}^{(0)}(\boldsymbol{\theta}, s)} \right] dN_i(s) \\ &\quad - \int_0^t \left[\begin{pmatrix} \widehat{W}_i(s) - \hat{\psi}_1(\gamma, s) \\ \boldsymbol{Z}_i(s) \end{pmatrix} - \frac{\breve{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}, \hat{\boldsymbol{\Psi}}, s)}{\breve{E}^{(0)}(\boldsymbol{\theta}, s)} \right] \breve{E}_i^{(0)}(\boldsymbol{\theta}, s)\breve{\lambda}_0(s)ds.\end{aligned}$$

A consistent estimate for $\boldsymbol{R}$ is given by

$$\breve{\boldsymbol{R}} = \breve{\boldsymbol{\mathcal{I}}}(\breve{\boldsymbol{\theta}}, \hat{\boldsymbol{\Psi}}, \tau)^{-1} \left[\frac{1}{n} \sum_i \breve{\boldsymbol{\Phi}}_i(\breve{\boldsymbol{\theta}}, \hat{\boldsymbol{\Psi}}, \tau)^{\otimes 2} + \sum_{i=2}^{n} \breve{\boldsymbol{\Phi}}_i(\breve{\boldsymbol{\theta}}, \hat{\boldsymbol{\Psi}}, \tau) \otimes \breve{\boldsymbol{\Phi}}_1(\breve{\boldsymbol{\theta}}, \hat{\boldsymbol{\Psi}}, \tau) \right] \breve{\boldsymbol{\mathcal{I}}}(\breve{\boldsymbol{\theta}}, \hat{\boldsymbol{\Psi}}, \tau)^{-1}.$$

4.4 Constructing the three groups of replicated longitudinal measurements

The flexibility of the above method is that it makes no distribution assumption on random effects and random errors. But there is a potential drawback that for each subject it requires three extra longitudinal data sets, $\{(\widetilde{W}_{ij,r}, t_{ij,r}), j = 1, \cdots, M\}, r = 1, 2, 3$.

Note that although in practice the replicated longitudinal observations $(\widetilde{W}_{ij,r}, t_{ij,r})$, $j = 1, \cdots, M, r = 1, 2, 3$ may not be available directly, we can construct replicated data sets in the following way. We may choose $M = q + 2$ or $M = q + 3$ and then select $3M$ longitudinal measurements from each individual if it has no less than $3M$ longitudinal observations. The $3M$ longitudinal measurements will be partitioned randomly into three groups as $(\widetilde{W}_{ij,r}, t_{ij,r}), j = 1, \cdots, M, r = 1, 2, 3$, which can be viewed as replicated longitudinal observations. These measurements will be used to calculate the consistent estimator $\hat{\boldsymbol{\Psi}}$. Then we keep the first group of longitudinal measurements $(\widetilde{W}_{ij,1}, t_{ij,1}), j = 1, \cdots, M$ unchanged and the rest of longitudinal observations for subject i are denoted as $(\widetilde{W}_{ij}, t_{ij}), j = 1, \cdots, m_i$.

Obviously, the larger value of M the smaller variance for $\xi_{i,r}(s)$. Thus a larger value of M leads to smaller variances of $\hat{\boldsymbol{\Psi}}$ and also the estimating equation (15). We expect that a larger value of M may result in a better estimator of $\boldsymbol{\theta}$. On the other hand, when using the above method to construct the three replicated data sets, subjects with less than $3M$ longitudinal observations are ignored. If we choose a very large value of M, then too many subjects will not be taken into account when estimating $\hat{\boldsymbol{\Psi}}$. This may lead to a larger variance of $\hat{\boldsymbol{\Psi}}$ and further a poor estimator of $\boldsymbol{\theta}$. In summary, we should choose M as large as possible, conditioning on that most subjects have at least $3M$ longitudinal measurements. Effects on the parameter estimators by choosing different values of M are discussed in the following section through simulation studies.

5 Simulation studies and data analysis

5.1 Simulation studies

We consider simulation scenarios in Tisiatis and Davidian (2001) where for simplicity there is a single time-dependent covariate $W_i(s)$ and no time-independent covariates are involved in the proportional hazard model. We choose a modified version of the scenarios in Tisiatis and Davidian (2001).

We assume $W_i(s) = \omega_{i0} + \omega_{i1}s$. Two different distributions for $(\omega_{i0}, \omega_{i1})$ are considered. They are (i) $(\omega_{i0}, \omega_{i1})$ is from a bivariate normal distribution with mean $(3.173, -0.0103)$ and covariance matrix $\boldsymbol{D}$ with elements $\boldsymbol{D} = (D_{11}, D_{12}, D_{22}) = (1.24, 0.039, 0.003)$; (ii) $(\omega_{i0}, \omega_{i1})$ follows a mixture of bivariate normal distribution, with mixing proportion 0.5 and mixture component $N(\boldsymbol{\mu}_k, \boldsymbol{D}_k), k = 1, 2$, where $\boldsymbol{\mu}_1 = (6.173, -0.0103)^T$, $\boldsymbol{\mu}_2 = (2.173, -0.0103)^T$ and $\boldsymbol{D}_1 = \boldsymbol{D}_2 = \boldsymbol{D}$. The maximum number of longitudinal observations for each subject is 24 and nominal times of observation for $W_i(s)$ are $\boldsymbol{t}_i = \{8 + 3j, j = 0, \cdots, 23\}$. Survival times are generated from the model $\lambda_i(s) = \exp(\gamma W_i(s))$ with $\gamma = -1$. The censoring distribution is exponential with mean 150 and with additional censoring at 80. We also consider two scenarios for the distribution of error terms: (a) a normally distributed error with distribution $N(0, 0.5)$; (b) the error term has a mixture normal distribution, $0.7N(-0.7, 0.01) + 0.3N(1.633, 0.01)$.

In each scenario sample sizes are chosen to be $n = 200$ and then 500 Monte Carlo data sets were generated. The parameter γ was estimated using four different methods: (1) using the 'ideal' estimator that can be obtained by fitting by partial likelihood with true values of $W_i(s)$; (2) using the conditional score (CS) estimator; (3) using the simple working likelihood (SWL) method; (4) using the generalized working likelihood (GWL) method with $M = 4, 5$. Other methods such as naive regression or method of 'last value carried forward' are not considered in the simulation studies since they are much less efficient than the conditional score estimator (Tisiatis and Davidian, 2001; Huang and Wang, 2000).

Table 1 is about here.

When the error term is normally distributed, from Table 1 we can see that the CS estimator and the SWL estimator work as well as the 'ideal' estimator in terms of the bias. The GWL estimators with $M = 4, 5$ also have very small bias. The GWL estimators have larger standard error estimates than the other two estimators, as the GWL method does not make the normal assumption for random error term.

When the error term is not normally distributed, results are summarized in Table 2.

Table 2 is about here.

We can see that the bias of CS estimator and SWL estimator increase since they are valid only for normal random errors, but the GWL estimators do not change much and actually they are very stable. We conclude that when random errors are not normally distributed, the GWL estimators with $M = 4, 5$ have much smaller bias than SWL and CS estimators. We also investigated other choices for σ^2 and obtained similar results. As the variance of ϵ_{ij} increases, the standard errors of all estimators increase. This is because a large variance of ϵ_{ij} results in a large variance for estimating equations and further leads to a large standard error for our estimator.

Tisiatis and Davidian (2001) pointed out that the estimating equation of conditional score method may have multi-roots. We then investigated the multi-roots problem for estimating equations of both methods. The typical score plots are shown in Figure 1. We can see that all methods have a solution close to the truth. The generalized working likelihood estimating equations have a single root. The conditional score method and the simple working likelihood method have multiple solutions. As Tisiatis and Davidian (2001) suggested, we may choose the naive regression estimator as starting value to locate the *correct* estimator.

Figure 1 is about here.

We also find out that when using generalized working likelihood method and choosing $M = 4$, the score function may have an *outlier* solution. In Figure 2 the solution for $M = 4$ are outliers while the solution for $M = 5$ is the truth. The outliers will result in poor estimator for the standard error and the mean. Similarly as Song and Huang (2005), when all estimators exist and no outliers exist, the general working likelihood estimators are stable and have a small bias, regardless of the distributions of the random effects and the error terms.

Figure 2 is about here.

5.2 Data analysis

To demonstrate the proposed methods we use the primary biliary cirrhosis (PBC) data set collected by the Mayo Clinic from 1974 to 1984. The PBC is a chronic disease that can eventually destroy some of the bile ducts linking liver to gut. When the PBC damages bile ducts, bile can no longer flow through them. Instead it builds up in the liver, damaging the liver cells and causing inflammation and scarring. In the clinical trial, survival status and laboratory results (e.g., serum bilirubin) of 312 patients were recorded. In this clinical trail, 158 out of 312 patients took the drug D-penicillamine and the other patients are in the control group. Serum bilirubin are measured at irregularly time points, recorded until death or censoring. Among the 312 patients, the maximum number of repeated measurement is 16. More details of the trail study and the data set can be found in Ding and Wang (2008) and Fleming and Harrington (1991).

We take the biomarker serum bilirubin as the time-dependent covariate process $W_i(s)$ and the treatment type as the time-independent covariate Z_i in the Cox regression model

$\lambda_i(s) = \lambda_0(s)\exp(\gamma W_i(s) + \beta Z_i)$, where the longitudinal model is $W_i(s) = \omega_{i0} + \omega_{i1}s$. In our analysis we took logarithm transformation on the serum bilirubin values as the longitudinal measurements. Six typical patients' serum bilirubin values are plotted in Figure 3.

Figure 3 is about here.

We fit the model using the conditional score method, simple working-likelihood method and generalized working-likelihood method. The maximum number of longitudinal observations for each subject is 16 and there are 32 patients who have more than 12 measurements on serum bilirubin. We choose $M = 4$ when using the generalized working-likelihood method. This means that $3M = 12$ longitudinal measurements are selected and they are randomly partitioned into three groups as replicated observations to estimate $\mathbf{\Psi}$.

The results are provided in Table 3. The three approaches provide similar results. All three methods suggest that the coefficient estimate $\hat{\gamma}$ for the latent process of serum bilirubin is not significant. The coefficient estimator based on baseline serum bilirubin in Fleming and Harrington (1991) is 0.8, significantly unequal to 0. This suggests that the baseline serum bilirubin is a risk factor for survival times but the serum bilirubin at later times after treatment is not. Fleming and Harrington (1991) also studied the treatment effect of D-penicillamine to PBC and their result is that the treatment is not significant. From Table 3 we can see that all three methods give the same result that the coefficient estimator $\hat{\beta}$ for treatment Z_i is not significantly unequal to 0.

Table 3 is about here.

6 Discussion

We have proposed new methods for joint modelling of survival events and error-contaminated time-dependent covariates. The estimators are easily computed and their large sample properties are shown. We suggest using the generalized working likelihood method in practice, since it does not lose much efficiency when the random error is normally distributed and it has the smallest bias if the error term is not normally distributed.

When using the general working likelihood method, however, we need replicated observations. Even if replicated observations do not exists, we can construct replicates from the longitudinal observations using the method described in Section 4. When partitioning the $3M$ longitudinal measurements into three groups, it should be done randomly for each subject. Note that we cannot partition the $3M$ measurements into three groups such that $t_{ij,1} < t_{ik,2} < t_{il,3}$ for $j, k, l = 1, \cdots, M$. This is because otherwise $(\widetilde{W}_{ij,r}, t_{ij,r}, j = 1, \cdots, M)$ will not have the same distribution for different values of r and then large sample properties of the estimator cannot be guaranteed.

All the existing parametric or nonparametric correction methods assume that the observation times t_{ij} are non-informative. In practice we may observe error-contaminated longitudinal points collected at informative observation times (Liang et.al., 2009). For such problems the existing methods are not valid. This deserves as our future research in the field.

A Regular conditions

C.1 The time τ is such that $\int_0^\tau \lambda_0(s)ds < \infty$.

C.2 Let

$$\boldsymbol{E}^{(k)}(W,\boldsymbol{\theta},s) = n^{-1}\sum_i \boldsymbol{E}_i^{(k)}(W_i,\boldsymbol{\theta},s) := n^{-1}\sum_i \begin{pmatrix} W_i(s) \\ \boldsymbol{Z}_i(s) \end{pmatrix}^{\otimes k} E_i^{(0)}(W_i,\boldsymbol{\theta},s), \quad k=0,1,2.$$

There exists a neighborhood $\boldsymbol{\Theta}$ of $\boldsymbol{\theta}_0$ and, respectively, scalar, vector and matrix functions $e^{(0)}$, $\boldsymbol{e}^{(1)}$ and $\boldsymbol{e}^{(2)}$ defined on $\boldsymbol{\Theta} \times [0,\tau]$ such that,

$$\sup_{s\in[0,\tau],\boldsymbol{\theta}\in\boldsymbol{\Theta}} ||\boldsymbol{E}^{(k)}(W,\boldsymbol{\theta},s) - \boldsymbol{e}^{(k)}(\boldsymbol{\theta},s)|| \to 0$$

in probability as $n \to \infty$.

C.3 Let $\boldsymbol{v} = \boldsymbol{e}^{(2)}/e^{(0)} - [\boldsymbol{e}^{(1)}/e^{(0)}]^{\otimes 2}$. Then for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $0 \leq s \leq \tau$,

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\theta}} e^{(0)}(\boldsymbol{\theta},s) &= \boldsymbol{e}^{(1)}(\boldsymbol{\theta},s) \\ \frac{\partial^2}{\partial\boldsymbol{\theta}^2} e^{(0)}(\boldsymbol{\theta},s) &= \boldsymbol{e}^{(2)}(\boldsymbol{\theta},s). \end{aligned}$$

C.4 The matrix

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}_0,\tau) = \int \boldsymbol{v}(\boldsymbol{\theta}_0,s)e^{(0)}(\boldsymbol{\theta}_0,s)\lambda_0(s)ds$$

is positive definite.

B Proof of Theorem 3.3

Proof. If $\sqrt{n}\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, \tau)$ converges weakly to a normal distribution, using the first-order Taylor extension for $\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, t)$ we have $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \Rightarrow N(\mathbf{0}, \boldsymbol{R})$, where

$$\boldsymbol{R} = \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0, \tau)^{-1} \boldsymbol{\Sigma}_U(\boldsymbol{\theta}_0, \sigma^2, \tau) \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0, \tau)^{-1}$$

and $\boldsymbol{\Sigma}_U(\boldsymbol{\theta}_0, \sigma^2, \tau) = \lim_{n\to\infty} Var[\sqrt{n}\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, \tau)]$. So we only need to show the asymptotic property for $\sqrt{n}\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, \tau)$.

We can write

$$\begin{aligned}
\sqrt{n}\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, t) &= n^{-1/2} \sum_i \int_0^t \left[\begin{pmatrix} \widehat{W}_i(s) \\ \boldsymbol{Z}_i(s) \end{pmatrix} - \frac{\hat{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, s)}{\hat{E}^{(0)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, s)} \right] dN_i(s) \\
&\quad -n^{-1/2} \sum_i \int_0^t \left[\begin{pmatrix} \widehat{W}_i(s) - \hat{\sigma}^2 v_i(s)\gamma \\ \boldsymbol{Z}_i(s) \end{pmatrix} - \frac{\hat{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, s)}{\hat{E}^{(0)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, s)} \right] \hat{E}_i^{(0)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, s)\lambda_0(s)ds
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
\sqrt{n}\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, t) &= n^{-1/2} \sum_i \int_0^t \left[\begin{pmatrix} \widehat{W}_i(s) \\ \boldsymbol{Z}_i(s) \end{pmatrix} - \frac{\boldsymbol{e}^{(1)}(\boldsymbol{\theta}_0, s)}{e^{(0)}(\boldsymbol{\theta}_0, s)} \right] dN_i(s) \\
&\quad -n^{-1/2} \sum_i \int_0^t \left[\begin{pmatrix} \widehat{W}_i(s) - \hat{\sigma}^2 v_i(s)\gamma \\ \boldsymbol{Z}_i(s) \end{pmatrix} - \frac{\boldsymbol{e}^{(1)}(\boldsymbol{\theta}_0, s)}{e^{(0)}(\boldsymbol{\theta}_0, s)} \right] \hat{E}_i^{(0)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, s)\lambda_0(s)ds \\
&\quad +n^{-1/2} \sum_i \int_0^t \left(\frac{\boldsymbol{e}^{(1)}(\boldsymbol{\theta}_0, s)}{e^{(0)}(\boldsymbol{\theta}_0, s)} - \frac{\hat{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, s)}{\hat{E}^{(0)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, s)} \right) \left(dN_i(s) - \hat{E}_i^{(0)}(\boldsymbol{\theta}_0, \sigma^2, s)\lambda_0(s)ds \right) \\
&:= I - II + III.
\end{aligned}$$

We know that $III = o_p(1)$, since under regular conditions $\mathcal{E}[dN_i(s) - \hat{E}_i^{(0)}(\boldsymbol{\theta}_0, \sigma^2, s)\lambda_0(s)ds | W_i(s)] = 0$ and $\sup_{\boldsymbol{\theta},s} \left| \frac{\boldsymbol{e}^{(1)}(\boldsymbol{\theta},s)}{e^{(0)}(\boldsymbol{\theta},s)} - \frac{\hat{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta},s)}{\hat{E}^{(0)}(\boldsymbol{\theta},s)} \right| = o(1)$. Thus we can write

$$\sqrt{n}\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, t) \;=\; I - II + o_p(1) = n^{-1/2} \sum_i \boldsymbol{\Phi}_i(\boldsymbol{\theta}_0, \hat{\sigma}^2, t) + o_p(1)$$

where $\boldsymbol{\Phi}_i(\boldsymbol{\theta}_0, \sigma^2, t)$ is

$$\int_0^t \left[\begin{pmatrix} \widehat{W}_i(s) \\ \boldsymbol{Z}_i(s) \end{pmatrix} - \frac{\boldsymbol{e}^{(1)}(\boldsymbol{\theta}_0, s)}{e^{(0)}(\boldsymbol{\theta}_0, s)} \right] dN_i(s) - \int_0^t \left[\begin{pmatrix} \widehat{W}_i(s) - \sigma^2 v_i(s)\gamma \\ \boldsymbol{Z}_i(s) \end{pmatrix} - \frac{\boldsymbol{e}^{(1)}(\boldsymbol{\theta}_0, s)}{e^{(0)}(\boldsymbol{\theta}_0, s)} \right] \hat{E}_i^{(0)}(\boldsymbol{\theta}_0, \sigma^2, s)\lambda_0(s)ds.$$

Since $\boldsymbol{\Phi}_i(\boldsymbol{\theta}_0, \sigma^2, t), i = 1, \cdots, n$ are i.i.d. random variables, we know that $n^{-1/2} \sum_i \boldsymbol{\Phi}_i(\boldsymbol{\theta}_0, \sigma^2, t)$ and $n^{-1/2} \sum_i [\boldsymbol{\Phi}_i(\boldsymbol{\theta}_0, \hat{\sigma}^2, t) - \boldsymbol{\Phi}_i(\boldsymbol{\theta}_0, \sigma^2, t)]$ both converge weakly to a Gaussian process. Therefore we have $\sqrt{n}\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, t)$ converges weakly to a normal distribution with mean $\mathbf{0}$ and covariance matrix

$$\boldsymbol{\Sigma}_U(\boldsymbol{\theta}_0, \sigma^2, t) = \lim_{n\to\infty} \left[\frac{1}{n} \sum_{i=1}^n \boldsymbol{\Phi}_i(\boldsymbol{\theta}_0, \hat{\sigma}^2, t)^{\otimes 2} + \sum_{i=2}^n \boldsymbol{\Phi}_i(\boldsymbol{\theta}_0, \hat{\sigma}^2, t) \otimes \boldsymbol{\Phi}_1(\boldsymbol{\theta}_0, \hat{\sigma}^2, t) \right]$$

According to the definition of $\hat{\boldsymbol{\Phi}}_i(\boldsymbol{\theta}_0, \hat{\sigma}^2, t)$ given in Theorem 3.3, we know that a consistent estimator for $\boldsymbol{\Sigma}_U(\boldsymbol{\theta}_0, \sigma^2, t)$ is

$$\hat{\boldsymbol{\Sigma}}_U(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2, t) = \frac{1}{n} \sum_i^n \hat{\boldsymbol{\Phi}}_i(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2, t)^{\otimes 2} + \sum_{i=2}^n \hat{\boldsymbol{\Phi}}_i(\boldsymbol{\theta}_0, \hat{\sigma}^2, t) \otimes \hat{\boldsymbol{\Phi}}_1(\boldsymbol{\theta}_0, \hat{\sigma}^2, t).$$

Thus a consistent estimator for $\boldsymbol{R}$ is $\hat{\boldsymbol{R}} = \hat{\boldsymbol{\mathcal{I}}}(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2, \tau)^{-1}\hat{\boldsymbol{\Sigma}}_U(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2, \tau)\hat{\boldsymbol{\mathcal{I}}}(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2, \tau)^{-1}$. □

C Proof of Theorem 4.1

The theorem follows from the obvious results

$$\frac{\varphi^{(1)}(s)}{\varphi^{(0)}(s)} = \frac{\mathcal{E}[\xi_{i,1}(s)\exp(\gamma\xi_{i,1}(s))]}{\mathcal{E}[\exp(\gamma\xi_{i,1}(s))]} = \frac{\mathcal{E}[(\xi_{i,1}(s) - \xi_{i,2}(s))\exp(\gamma\xi_{i,1}(s) - \gamma\xi_{i,3}(s))]}{\mathcal{E}[\exp(\gamma\xi_{i,1}(s) - \gamma\xi_{i,3}(s))]},$$

$$\frac{\varphi^{(2)}(s)}{\varphi^{(0)}(s)} = \frac{\mathcal{E}[\xi_{i,1}(s)^2 \exp(\gamma\xi_{i,1}(s))]}{\mathcal{E}[\exp(\gamma\xi_{i,1}(s))]} = \frac{\mathcal{E}[\{(\xi_{i,1}(s) - \xi_{i,2}(s))^2 - \xi_{i,2}(s)^2\} \exp(\gamma\xi_{i,1}(s) - \gamma\xi_{i,3}(s))]}{\mathcal{E}[\exp(\gamma\xi_{i,1}(s) - \gamma\xi_{i,3}(s))]}$$

and $\mathcal{E}[\xi_{i,2}(s)^2] = \sigma^2 \mathcal{E}\left[\boldsymbol{s}^T[\boldsymbol{A}_{i,2}^T \boldsymbol{A}_{i,2}]^{-1}\boldsymbol{s}\right]$.

References

Cox D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, B, **34**:187-220.

Dafni U. G. and Tsiatis A. A. (1998). Evaluating surrogate markers of clinical outcome measured with error. *Biometrics*, **54**:1445-1462.

DeGruttola V. and Tu X. M. (1994). Modeling progression of CD-4 lymphocyte count and its relationship to survival time. *Biometrics*, **50**:1003-1014.

Faucett C. J. and Thomas D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Statstics in Medicine*, **1**:465-480.

Fleming T. R. and Harrington D. P. (1991). *Counting Processes and Survival Analysis*, John Wiley & Sons, Inc.

Henderson R., Diggle P. and Dobson A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **1**:465-480.

Hu P., Tsiatis A. A. and Davidian M. (1998). Estimating the parameters in the Cox model when covariate variables are measured with error. *Biometrics*, **54**:1407-1419.

Huang Y. and Wang C. Y. (2000). Cox regression with accurate covariates unascertainable: A nonparametric-correction approach. *Journal of the American Statistical Association*, **95**:1209-1219.

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

Nakamura T. (1990). Corrected score function for errors in variables models: methodology and application to generalized linear models. *Biometrika*, **77**:127-137.

Prentice R. (1982). Covariate measurement errors and parameter estimates in a failure time regression measurement error models. *Biometrika*, **74**:703-716.

Song X., Davidian M. and Tsiatis A. A. (2002a). An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. *Biostatistics*, **3**:511-528.

Song X., Davidian M. and Tsiatis A. A. (2002b). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, **88**:447-458.

Jimin Ding and Jane-Ling Wang. (2008). Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics*, **64**:546-556.

Tisiatis A. A. and Davidian M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, **88**:447-458.

Tisiatis A. A. and DeGruttola V. and Wulfsohn M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error: Applications to survival and CD4 counts in patients with AIDS. *J. Am. Statist. Assoc.*, **90**:27-37.

Song X. and Huang Y. (2005). On corrected score approach for proportional hazards model with covariate measurement error. *Biometrics*, **61**:702-714.

Wang C. Y. and Hsu L. and Feng Z. D. and Prentice R. L. (1997). Regression Calibration in Failure Time Regression. *Biometrics*, **53**:131-145.

Wang C. Y. (2006). Corrected score estimator for joint modelling of longitudinal and failure time data. *Statistica Sinica*, **16**:235-253.

Wulfson M. S. and Tsiatis A. A. (1997). A joint model for survival and longitudianl data measured with error. *Biometrics*, **53**:330-339.

Table 1: Simulation results for two underlying random effect distributions, when error term is $N(0, 0.5)$. I, 'ideal'; CS, conditional score estimator; SWL, simple working likelihood estimator; GWL, Generalized working likelihood estimator with $M = 4, 5$; SD, Monte Carlo standard deviation; SE, average of estimated standard errors.

	$n = 200$					
	Normal covariate			Mixture covariate		
Method	Mean	SD	SE	Mean	SD	SE
I	-0.9978	0.087	0.086	-0.9910	0.083	0.078
CS	-0.9888	0.127	0.118	-1.0027	0.129	0.112
SWL	-0.9971	0.143	0.140	-1.0103	0.140	0.129
GWL,4	-0.9957	0.249	0.268	-1.0553	0.344	0.382
GWL,5	-0.9852	0.190	0.193	-1.0243	0.217	0.227

Table 2: Simulation results for two underlying random effect distributions, when error term is mixed normal as $0.7N(-0.7, 0.01) + 0.3N(1.633, 0.01)$. I, 'ideal'; CS, conditional score estimator; SWL, simple working likelihood estimator; GWL, Generalized working likelihood estimator with $M = 4, 5$; SD, Monte Carlo standard deviation; SE, average of estimated standard errors.

	$n = 200$					
	Normal covariate			Mixture covariate		
Method	Mean	SD	SE	Mean	SD	SE
I	-0.9978	0.087	0.086	-0.9910	0.083	0.078
CS	-1.0808	0.135	0.124	-1.0810	0.132	0.119
SWL	-1.0938	0.236	0.273	-1.0787	0.212	0.209
GWL4	-1.0263	0.419	0.490	-1.0269	0.403	0.431
GWL5	-0.9794	0.246	0.234	-0.9937	0.244	0.247

Table 3: Results for the PBC data.

	$\hat{\gamma}$ (sd)	$\hat{\beta}$ (sd)
CS	-0.029 (0.069)	0.053 (0.265)
SWL	-0.033 (0.055)	0.054 (0.160)
GWL	-0.004 (0.099)	0.077 (0.248)

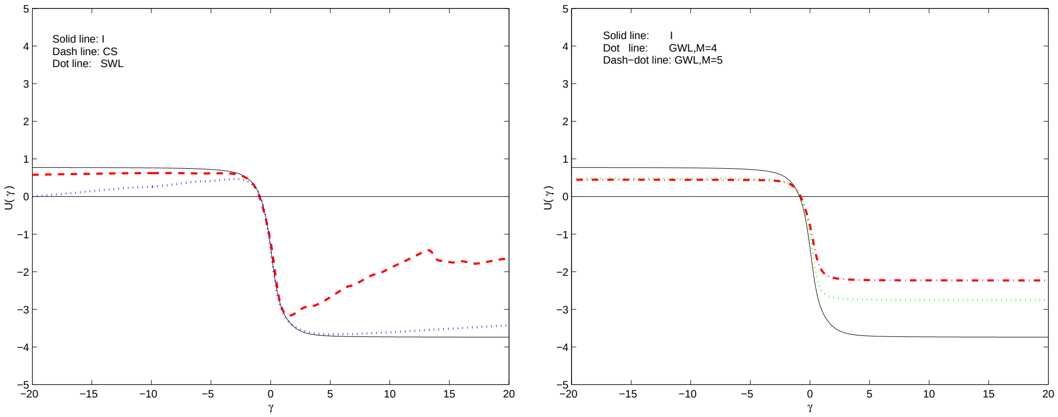

Figure 1: Typical score plots for simulation data sets. I, 'ideal' method; CS, conditional score method; SWL, simple working likelihood method; GWL, generalized working likelihood method.

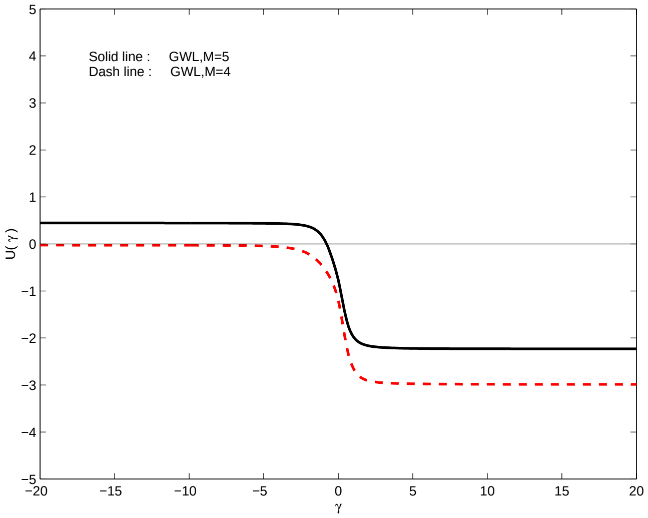

Figure 2: Typical score plots for a simulation data set. GWL, generalized working likelihood method.

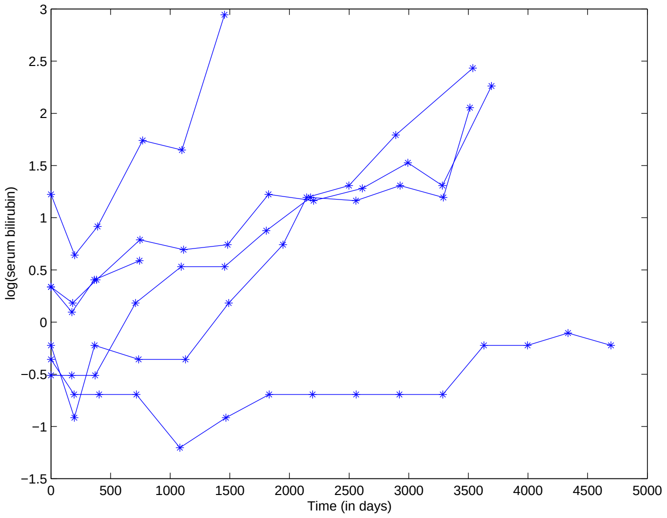

Figure 3: Longitudinal observation plot for six patients.