Modelling survival events with longitudinal covariates measured with error

Hongsheng Dai,

School of CEM, Watts Building, University of Brighton, Brighton, BN2, 4GJ UK

E–mail: h.dai@brighton.ac.uk;

Jianxin Pan, University of Manchester, UK;

Yanchun Bao, Yunnan Normal University, China

September 6, 2011

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 depends on the underlying time-dependent covariate 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 any parametric distribution assumption on random effects and random errors. Asymptotic properties for parameter estimators are provided. Simulation results show that under certain situations the proposed methods are more efficient than the existing methods.

Key words: Longitudinal measurements; Partial likelihood; Linear mixed model; Random effects; Proportional hazard model.

1 Introduction

This research work is to develop new methods for the Cox proportional hazard model (Cox, 1972) with error-contaminated time-dependent covariates, avoiding the parametric assumptions in existing methods. The Cox model, which is widely used to study the relationship between survival events and time-dependent or time-independent covariates, assumes that the failure time hazard rate function $\lambda(s)$ relates to a latent process $W(s)$, an observed process $\boldsymbol{Z}(s)$ and baseline hazard $\lambda_0(s)$ through $\lambda(s) = \lambda_0(s)\exp(\gamma W(s) + \boldsymbol{\beta}^T \boldsymbol{Z}(s))$, where γ and β are the unknown parameters. To implement the Cox model most existing methods require the latent time-dependent process $W(s)$ to be fully observed. In practice, however, $W(s)$ is measured intermittently and 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 leads to 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 more complex polynomial models. 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 research works (DeGruttola and Tu, 1994; Faucett and Thomas, 1996; Henderson et al., 2000; Wulfson and Tsiatis, 1997) assumed that the random effects and random error follow Gaussian distributions and they used EM algorithms or Bayesian approaches to deal with the unobserved latent process $W(s)$. However, the normal distribution assumption on random effects and random errors may not be true in practice. Misspecification of the distributions of random effects or random errors will result in biased estimates (Tsiatis and Davidian, 2001; Song and Huang, 2005; Wang, 2006).

Hu et al. (1998) relaxed the normal assumption by assuming that the density of underlying covariates belongs to a smooth class. These approaches involve intensive EM-algorithm computation. Huang and Wang (2000) and Song and Huang (2005) proposed a corrected score approach which does not require any distribution assumption on 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 observations for each subject. In many applications, however, the underlying covariates are time-dependent and replicated observations for each subject may not be available. Recently, another interesting method is proposed by Wang (2006). With the normal distribution assumption for error terms, Wang's method does not put any distribution restriction 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 estimator, the conditional score estimator, is proposed by Tsiatis and Davidian (2001). When the hazard rate $\lambda(s)$ depends on the time-dependent underlying process and the error term is normally distributed, conditional score (CS) estimator does not require any distribution assumption on $W(s)$.

In this paper, we consider the general case where the failure time hazard rate depends on a latent time-dependent process. With the normal distribution assumption on ϵ, a simple working-likelihood (SWL) estimator is proposed without any distribution assumption on ω_0 and ω_1. The simple working-likelihood estimator is proved to be consistent and asymptotically normal under some regularity conditions. A consistent covariance estimator is also provided. Then we relax the normal distribution assumption on ϵ. A generalized working-likelihood (GWL) estimator is introduced for such cases. Consistency and asymptotic distribution of the GWL estimator are also shown. Simulation studies demonstrate that when the error term ϵ follows a normal distribution, the SWL estimator has smaller bias but larger variance than the CS estimator. Numerical studies also show that the GWL estimator works well and it has much smaller bias than SWL and CS estimators when ϵ and (ω_0, ω_1) are both non-normally distributed. This paper is organized as follows. Models and notations are given in Section 2. In Section 3 we propose the simple working-likelihood estimator. The generalized working-likelihood estimator is provided in Section 4. Numerical studies and discussions are given in sections 5 and 6.

2 Models and notations

Let $\boldsymbol{Z}_i(s)$ be the observed time-dependent process and $W_i(s)$ be the unobserved time-dependent process for the ith subject, $i = 1, \cdots, n$. 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. We also assume that $W_i(s) = \sum_{j=0}^{q-1} \omega_{ij} s^j$. Here

$\boldsymbol{\omega}_i = (\omega_{i0}, \cdots, \omega_{i,q-1})^T$ is the random effect for the ith subject. 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

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

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 $\delta_i = I(T_i \leq C_i)$, where the censoring variable C_i is independent of T_i. There is 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 $\boldsymbol{\epsilon}_i$. The Cox proportional hazard model assumes that the hazard rate is a function of covariates through the following form,

$$\lambda_i(s) = \lambda_0(s) \exp(\gamma W_i(s) + \boldsymbol{\beta}^T \boldsymbol{Z}_i(s)) \tag{2}$$

where $\lambda_0(s)$ is an arbitrary baseline hazard function. Our aim is to estimate the unknown parameters $\boldsymbol{\beta}$ and γ in the above model.

Define counting process $dN_i(s) = I[s \leq \tilde{T}_i \leq s + ds, \delta_i = 1, t_{iq} \leq s]$ and at-risk process $Y_i(s) = I[\tilde{T}_i \geq s, t_{iq} \leq s]$. The log-partial likelihood function for parameter $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta}^T)^T$ is

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

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 $\boldsymbol{E}^{(1)}(W,\boldsymbol{\theta},s)$ is the partial derivative of $\partial E^{(0)}(W,\boldsymbol{\theta},s)$ with respect to the parameters $(\gamma,\boldsymbol{\beta}^T)^T$, given by

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

Using martingale theories and under some regularity conditions, it is straightforward to show that the maximum likelihood estimate based on (3) is consistent. Details can be found in Fleming and Harrington (1991).

When $W_i(s)$ is measured with error and intermittently, the score functions in (4) are not available. To solve this problem, a naive approach is to estimate $W_i(s)$ using the Least Squares estimates (LSE) and then to apply these estimates as the true values. Another method is to use the regression calibration estimator, which is to replace $W_i(s)$ with its conditional expectation given the longitudinal measurements. These approaches, however, introduce a severe bias to the estimate of γ. Detailed discussions and comparisons can be found in Tsiatis and Davidian (2001) and Wang (2006). Based on a sufficient statistic for $W_i(s)$ Tsiatis and Davidian (2001) proposed a conditional score estimator, with no distribution assumption given to the random effect $\boldsymbol{\omega}_i$. The conditional score estimator is more efficient than the naive approach and regression calibration. We will briefly compare the conditional score estimator with the proposed estimator in the following section.

3 A simple working likelihood estimator

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

The score function (4) is not available since $W_i(s)$ is measured intermittently and with error. To derive a working score function it is natural to consider using the least squares estimate (LSE) of $W_i(s)$. 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. Let $\boldsymbol{s} = (1, s, \cdots, s^{q-1})^T$ 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 $\mathrm{Var}(\widehat{W}_i(s)|W_i(s)) = \sigma^2 v_i(s)$ where $v_i(s) := \boldsymbol{s}^T(\boldsymbol{A}_i^T \boldsymbol{A}_i)^{-1}\boldsymbol{s}$. A consistent estimator (Tsiatis and Davidian, 2001) for σ^2 is $\hat{\sigma}^2 = \dfrac{\sum_i I[m_i > q] R_i}{\sum_i I[m_i > q](m_i - q)}$, where R_i is the residual sum of squares for subject i for the least squares fit to all the m_i observations. If we simply replace $W_i(t)$ in (4) by $\hat{W}_i(t)$, we will obtain a biased estimate (Tsiatis and Davidian, 2001). To obtain an unbiased working score function we consider the following procedure.

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

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

We know 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 $\mathcal{E}\{\exp[\gamma \widehat{W}_i(s)]|W_i(s)\} = \exp[\gamma W_i(s) +$

$\gamma^2\sigma^2 v_i(s)/2]$, where $\mathcal{E}$ means 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 regularity conditions (see Appendix A) and according to (5) we have that $\hat{E}^{(0)}(\boldsymbol{\theta}, \sigma^2, s)$ is asymptotically equivalent to $E^{(0)}(W, \boldsymbol{\theta}, s)$. Therefore if σ^2 is known the working likelihood function,

$$\hat{l}_n(\boldsymbol{\theta}, \sigma^2) = n^{-1} \sum_i \int \left[\gamma \widehat{W}_i(s) + \boldsymbol{\beta}^T \boldsymbol{Z}_i(s) - \log \hat{E}^{(0)}(\boldsymbol{\theta}, \sigma^2, s)\right] dN_i(s), \tag{6}$$

is asymptotically unbiased for $l(\boldsymbol{\theta})$ given in (3). To find the MLE based on (6), we take derivatives over γ and $\boldsymbol{\beta}$ and obtain the following score functions

$$\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}, \sigma^2, \tau) \quad := \quad n^{-1} \sum_i \int_0^\tau \left[\begin{pmatrix} \widehat{W}_i(s) \\ \boldsymbol{Z}_i(s) \end{pmatrix} - \frac{\hat{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}, \sigma^2, s)}{\hat{E}^{(0)}(\boldsymbol{\theta}, \sigma^2, s)} \right] dN_i(s) = \boldsymbol{0} \tag{7}$$

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

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

If we replace σ^2 by $\hat{\sigma}^2$ in (7), we will have that $\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}, \hat{\sigma}^2, \tau)$ is also an unbiased score function. This is given by the following theorem.

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 regularity conditions given in Appendix A, we have $\lim_{n\to\infty} \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$ in probability. We also have $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \Rightarrow N(\boldsymbol{0}, \boldsymbol{R}_1)$, where $\boldsymbol{R}_1$ is given in Appendix B.*

See Appendix B for a sketch of proof. An estimator for $\boldsymbol{R}_1$ is also given in Appendix B. Let $\bar{N}(s) = \sum_i N_i(s)/n$. A consistent estimator for $\lambda_0(s)ds$ is $\hat{\lambda}_0(s)ds = d\bar{N}(s)/\hat{E}^{(0)}(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2, s)$.

Note that the conditional score estimator in Tsiatis and Davidian (2001) used a similar estimating equation as (7) but replaced $\hat{E}_i^{(0)}(\boldsymbol{\theta}, \sigma^2, s)$ with $\hat{E}_{i*}^{(0)}(W_i, \boldsymbol{\theta}, s) := \exp[\gamma(\hat{W}_i(s) + \gamma\sigma^2 v_i(s)dN_i(s)) - \gamma^2\sigma^2 v_i(s)/2 + \boldsymbol{\beta}^T\boldsymbol{Z}_i(s)]Y_i(s)$. If we define the filtration $\mathcal{F}_s = \bigvee_i \mathcal{F}_{i,s}$ with $\mathcal{F}_{i,s} = \sigma\{I[\tilde{T}_i \leq u], \delta_i, \boldsymbol{\omega}_i, \boldsymbol{Z}_i(u), u \leq s\}$, the process $\hat{E}_{i*}^{(0)}(W_i, \boldsymbol{\theta}, s)$ is not predictable with respect to $\mathcal{F}_s$ since it is a function of $dN_i(s)$. This might make it difficult to use martingale theories to prove the asymptotic normality for the estimator. Tsiatis and Davidian (2001) only provided a heuristic proof. The proposed SWL estimator based on the estimating equation (7), however, does not involve $dN_i(s)$. This means that the integrand in (7) will be predictable with respect to $\mathcal{F}_s$. This enables us to show asymptotic normality with martingale theories.

4 A general working likelihood estimator

In the previous section we proposed a simple working likelihood estimator based on the assumption that ϵ is normally distributed. This assumption may not be true in practice. Throughout this section, we relax the normal distribution assumption for $\boldsymbol{\epsilon}_i$ and only assume $\epsilon_{ij}, j = 1, \cdots, m_i, i = 1, \cdots, n$ are i.i.d. with mean 0 and finite variance σ^2. 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

First we need to derive an unbiased working likelihood function. Since we can calculated the LSE $\widehat{W}_i(s)$ but do not know $W_i(s)$, it is necessary to start from studying the properties of their difference, $\xi_i(s) = \widehat{W}_i(s) - W_i(s)$. We have

$$\xi_i(s) = \boldsymbol{s}^T(\boldsymbol{A}_i^T\boldsymbol{A}_i)^{-1}\boldsymbol{A}_i^T\boldsymbol{\epsilon}_i. \tag{8}$$

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 the same.

Obviously $\xi_i(s)$ cannot be calculated since we do not know $W_i(s)$. But we can bypass this if 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 estimate 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 (8) and each subject has the same number of replicated longitudinal measurements, we know that $\xi_{i,1}(s), i = 1, \cdots, n$ are i.i.d. random variables. We cannot obtain the values of $\xi_i(s)$ and $\xi_{i,1}(s)$, but we can have their difference $\xi_i(s) - \xi_{i,1}(s) = \widehat{W}_i(s) - \widehat{W}_{i,1}(s)$. The unbiased working likelihood function can be derived based on this difference. The method is as follows.

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

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

Note that $\breve{E}_i^{(0)}(\boldsymbol{\theta}, s)$ is $E_i^{(0)}(W_i, \boldsymbol{\theta}, s)$ with W_i replaced by $\widehat{W}_{i,1}$, the Least Squares estimate 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 $\lim_{n\to\infty} \breve{E}^{(0)}(\boldsymbol{\theta}, s)/\varphi^{(0)}(\gamma, s) = \lim_{n\to\infty} E^{(0)}(\boldsymbol{\theta}, s) = e^{(0)}(\boldsymbol{\theta}, s)$ under some regularity conditions. Similarly as the results in Section 3, if $\varphi^{(0)}(\gamma, s)$ is known we have an unbiased log-partial likelihood function as

$$\breve{l}_n(\boldsymbol{\theta}) = n^{-1} \sum_i \int \left[\gamma \widehat{W}_i(s) + \boldsymbol{\beta}^T \boldsymbol{Z}_i(s) - \log \frac{\breve{E}^{(0)}(\boldsymbol{\theta}, s)}{\varphi^{(0)}(\gamma, s)} \right] dN_i(s). \tag{9}$$

Define

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

Then unbiased estimating equations are

$$\breve{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}, \boldsymbol{\Psi}, \tau) \quad := \quad n^{-1} \sum_i \int_0^\tau \left[\begin{pmatrix} \widehat{W}_i(s) \\ \boldsymbol{Z}_i(s) \end{pmatrix} - \frac{\breve{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}, \boldsymbol{\Psi}, s)}{\breve{E}^{(0)}(\boldsymbol{\theta}, s)} \right] dN_i(s) = \boldsymbol{0}, \tag{10}$$

where

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

Note that if we replace $\boldsymbol{\Psi} = \{\psi_1(\gamma, s), \psi_2(\gamma, s)\}$ in (10) by its consistent estimator $\hat{\Psi}$, then we can calculate the MLE by solving score functions $\breve{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}, \hat{\boldsymbol{\Psi}}, \tau) = \boldsymbol{0}$.

Note that Song and Huang (2005) provided a similar approach for error contaminated time-independent covariates. Therefore this method can be viewed as an extension of their method to error contaminated time-dependent processes.

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

To obtain an MLE by solving score functions $\breve{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}, \hat{\boldsymbol{\Psi}}, \tau) = \boldsymbol{0}$, we need to find a consistent estimator $\hat{\boldsymbol{\Psi}}$. Suppose that for each subject, besides the M extra longitudinal measurements $(\widetilde{W}_{ij,1}, t_{ij,1}), j = 1, \cdots, M$, there are another 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 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 $\hat{\boldsymbol{\Psi}} = (\hat{\psi}_1(\gamma, s), \hat{\psi}_2(\gamma, s))$ given in the above theorem, we have the unbiased estimating equations as

$$\breve{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}, \hat{\boldsymbol{\Psi}}, \tau) \;:=\; n^{-1} \sum_i \int_0^\tau \left[\begin{pmatrix} \widehat{W}_i(s) \\ \boldsymbol{Z}_i(s) \end{pmatrix} - \frac{\breve{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}, \hat{\boldsymbol{\Psi}}, s)}{\breve{E}^{(0)}(\boldsymbol{\theta}, s)} \right] dN_i(s) = \boldsymbol{0}. \qquad (11)$$

Similarly as the proof for Theorem 3.1 we can show that the estimated value $\breve{\boldsymbol{\theta}}$ by solving (11) is consistent. We can also show that $\sqrt{n}(\breve{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \Rightarrow N(\boldsymbol{0}, \boldsymbol{R}_2)$. An estimator of $\boldsymbol{R}_2$ is given in Appendix D.

4.3 Constructing the replicated longitudinal measurements

The advantage of the above method is that it makes no distribution assumption on random effects and random errors. But this also leads to its drawback that for each subject it needs 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 $(\widetilde{W}_{ij,r}, t_{ij,r})$, $j = 1, \cdots, M, r = 1, 2, 3$ are not available directly, we can achieve replicated data sets in the following way. We may choose $M = q + 2$ or $M = q + 3$ and 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 the longitudinal measurements $(\widetilde{W}_{ij,1}, t_{ij,1}), j = 1, \cdots, M$ unchanged and the rest of longitudinal observations for each subject are denoted as $(\widetilde{W}_{ij}, t_{ij}), j = 1, \cdots, m_i$.

The larger value of M, the smaller the variance for $\xi_{i,r}(s)$. Thus a larger value of M causes smaller variances of $\hat{\boldsymbol{\Psi}}$ and the estimating equation (11). Using the above method for constructing the three replicated data sets, subjects with less than $3M$ longitudinal observations have to be ignored when calculating $\hat{\boldsymbol{\Psi}}$. If we choose a very large value of M, then many subjects may not be taken into account for the calculation of $\hat{\boldsymbol{\Psi}}$. So provided that a sufficient number of subjects have no less than $3M$ longitudinal observations we will expect that a larger value of M leads to better estimates of $\boldsymbol{\theta}$. Effects on the parameter estimates of 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 Tsiatis and Davidian (2001) where for simplicity there is a single time-dependent process $W_i(s)$ and no time-independent covariates are involved in the proportional hazard model. We choose a modified version of the scenarios in Tsiatis 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 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 four scenarios for the distribution of error terms: (a) a normally distributed error with distribution $N(0, 0.5)$; (b) a normally distributed error with distribution $N(0, 1.5)$; (c) the error term has a mixture normal distribution, $0.7N(0.7, 0.01) + 0.3N(-1.633, 0.01)$; (d) the error term has a mixture normal distribution, $0.7N(0.7, 0.1) + 0.3N(-1.633, 0.1)$. Note that scenario (c) chooses a small variance 0.01 for each component but the variance of ϵ under scenario (c) is $0.7(0.7^2 + 0.01) + 0.3(1.633^2 + 0.01) = 1.15$ which is not small. Scenario (d) increases the component variance from 0.01 to 0.1 and the variance of ϵ changes from 1.15 to $0.7(0.7^2 + 0.1) + 0.3(1.633^2 + 0.1) = 1.24$.

In each scenario two sample sizes $n = 200$ and $n = 100$ are chosen and 2000 Monte Carlo data sets were generated. The parameter γ was estimated using four different methods: (1) using the 'ideal' estimator that could 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 (Tsiatis and Davidian, 2001; Huang and Wang, 2000).

Scenarios (a) and (b)

When the error term is normally distributed (scenarios (a) and (b)), from Table 1 we can

Table 1: Simulation results for two underlying random effect distributions, when error term is $N(0, \sigma^2)$. 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. 'Normal' means normally distributed random effects ω_i and 'mixture' means that ω_i follows a mixture normal distribution.

(a)	$n = 200, \sigma^2 = 0.5$					
		Normal			Mixture	
Method	Estimate (bias)	SD	SE	Estimate (bias)	SD	SE
I	-0.9988 (0.0012)	0.111	0.116	-0.9931 (0.0069)	0.106	0.117
CS	-0.9911 (0.0089)	0.119	0.133	-1.0056 (0.0056)	0.113	0.127
SWL	-1.0084 (0.0084)	0.148	0.145	-1.0129 (0.0129)	0.139	0.129
GWL,4	-1.0303 (0.0303)	0.279	0.285	-1.0506 (0.0506)	0.250	0.254
GWL,5	-1.0081 (0.0081)	0.220	0.205	-1.0135 (0.0135)	0.211	0.219
(a)	$n = 100, \sigma^2 = 0.5$					
		Normal			Mixture	
Method	Estimate (bias)	SD	SE	Estimate (bias)	SD	SE
I	-0.9988 (0.0012)	0.111	0.116	-0.9931 (0.0069)	0.106	0.117
CS	-1.0167 (0.0167)	0.178	0.201	-1.0394 (0.0394)	0.172	0.209
SWL	-1.0484 (0.0484)	0.247	0.259	-1.0346 (0.0346)	0.223	0.240
GWL,4	-0.9757 (0.0243)	0.278	0.293	-1.0173 (0.0173)	0.268	0.280
GWL,5	-0.9593 (0.0407)	0.227	0.235	-1.0242 (0.0242)	0.230	0.241
(b)	$n = 200, \sigma^2 = 1.5$					
		Normal			Mixture	
Method	Estimate (bias)	SD	SE	Estimate (bias)	SD	SE
I	-0.9988 (0.0012)	0.111	0.116	-0.9931(0.0069)	0.106	0.117
CS	-1.0094 (0.0094)	0.124	0.138	-0.9857 (0.0143)	0.123	0.138
SWL	-1.0111 (0.0111)	0.238	0.247	-0.9842 (0.0158)	0.275	0.286
GWL,4	-0.9717 (0.0283)	0.379	0.371	-0.9877 (0.0123)	0.355	0.377
GWL,5	-0.9820 (0.0180)	0.326	0.355	-0.9938 (0.0062)	0.341	0.357
(b)	$n = 100, \sigma^2 = 1.5$					
		Normal			Mixture	
Method	Estimate (bias)	SD	SE	Estimate (bias)	SD	SE
I	-0.9988 (0.0012)	0.111	0.116	-0.9931(0.0069)	0.106	0.117
CS	-0.9616 (0.0384)	0.193	0.240	-1.0489 (0.0489)	0.201	0.252
SWL	-1.0208 (0.0208)	0.309	0.339	-1.0508 (0.0508)	0.283	0.322
GWL,4	-0.9290 (0.0710)	0.379	0.391	-0.9779 (0.0221)	0.388	0.409
GWL,5	-0.9425 (0.0575)	0.365	0.389	-0.9935 (0.0065)	0.370	0.369

see that with $n = 200$ the CS estimator and the SWL estimator work as well as the 'ideal' estimator with respect to the bias, but the SWL estimator has larger standard error estimates than the CS estimator. When sample size $n = 200$ the bias of the GWL estimator with $M = 5$ is as small as the bias of the CS and SWL estimators. But the bias of the GWL estimator with $M = 4$ is larger than the other estimators. When the sample size $n = 100$, all estimators have a relatively larger bias than the results with $n = 200$. The GWL estimators have larger standard error estimates than the other two estimators. This is reasonable since the GWL method does not make the normal assumption for random error term. From Table 1 we can also see that if we increase the variance of ϵ from 0.5 to 1.5, the variances of the SWL and GWL estimators increase. This is because larger variance of ϵ results in larger variance for estimating equations and further lead to larger standard errors for our estimators.

Scenarios (c) and (d)

When the error term follows a mixture of normal distributions, results are summarized in Table 2. Comparing to Table 1, we can see that the bias of CS estimate and SWL estimate increase since they are valid only for normal random errors, but the GWL estimates did not change much and they are very stable. The CS and SWL estimators are biased estimators when ϵ is from a mixture of normal distributions. This can be seen from Table 2. When we increases the sample sizes from $n = 100$ to $n = 200$, the bias of the CS estimator and the SWL estimator increase. For example bias increases from 0.0480 to 0.0809 for CS estimator with normal random effects. On the contrary, the bias of the GWL estimator decreases when sample size increases from $n = 100$ to $n = 200$ which implies that the GWL estimator is consistent when ϵ is from a mixture of normal distributions.

We also notice that the bias of the GWL estimators may be large when sample size $n = 100$. For example the bias of the GWL estimator with $M = 4$ is 0.0933 with $\sigma_c^2 = 0.1$ and ω_i normally distributed. But the bias decreases a lot when sample size increases to $n = 200$.

Table 2: Simulation results for two underlying random effect distributions, when error term is mixed normal as $0.7N(0.7, \sigma_c^2) + 0.3N(-1.633, \sigma_c^2)$. 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. 'Normal' means normally distributed random effects ω_i and 'mixture' means that ω_i follows a mixture normal distribution.

(c)	$n = 200, \sigma_c^2 = 0.01$					
		Normal			Mixture	
Method	Estimate (bias)	SD	SE	Estimate (bias)	SD	SE
I	-0.9988 (0.0012)	0.111	0.116	-0.9931 (0.0069)	0.106	0.117
CS	-1.0809 (0.0809)	0.123	0.148	-1.0689 (0.0689)	0.110	0.130
SWL	-1.1157 (0.1157)	0.252	0.296	-1.0795 (0.0795)	0.204	0.231
GWL,4	-0.9709 (0.0291)	0.406	0.444	-1.0274 (0.0274)	0.366	0.423
GWL,5	-0.9811 (0.0189)	0.256	0.247	-1.0172 (0.0172)	0.289	0.288

(c)	$n = 100, \sigma_c^2 = 0.01$					
		Normal			Mixture	
Method	Estimate (bias)	SD	SE	Estimate (bias)	SD	SE
I	-0.9988 (0.0012)	0.111	0.116	-0.9931 (0.0069)	0.106	0.117
CS	-1.0480 (0.0480)	0.228	0.190	-1.0495 (0.0495)	0.184	0.225
SWL	-0.9694 (0.0306)	0.471	0.504	-0.9624 (0.0376)	0.391	0.438
GWL,4	-0.9503 (0.0493)	0.438	0.471	-0.9631 (0.0369)	0.377	0.392
GWL,5	-0.9881 (0.0119)	0.280	0.306	-0.9887 (0.0113)	0.276	0.282

(d)	$n = 200, \sigma_c^2 = 0.1$					
		Normal			Mixture	
Method	Estimate (bias)	SD	SE	Estimate (bias)	SD	SE
I	-0.9988 (0.0012)	0.111	0.116	-0.9931 (0.0069)	0.106	0.117
CS	-0.9470 (0.0530)	0.105	0.126	-0.9282 (0.0718)	0.105	0.140
SWL	-1.0684 (0.0684)	0.202	0.229	-1.0542 (0.0542)	0.194	0.205
GWL,4	-0.9859 (0.0141)	0.382	0.409	-1.0412 (0.0412)	0.371	0.413
GWL,5	-0.9931 (0.0069)	0.298	0.315	-1.0228 (0.0228)	0.309	0.318

(d)	$n = 100, \sigma_c^2 = 0.1$					
		Normal			Mixture	
Method	Estimate (bias)	SD	SE	Estimate (bias)	SD	SE
I	-0.9988 (0.0012)	0.111	0.116	-0.9931 (0.0069)	0.106	0.117
CS	-0.9127 (0.0873)	0.170	0.246	-1.0402 (0.0402)	0.235	0.184
SWL	-1.0342 (0.0342)	0.306	0.362	-1.0669 (0.0669)	0.267	0.271
GWL,4	-0.9067 (0.0933)	0.399	0.485	-0.9665 (0.0335)	0.404	0.483
GWL,5	-0.9448 (0.0542)	0.336	0.394	-0.9856 (0.0144)	0.303	0.339

Therefore we may conclude that for small sample size the GWL estimators may not work well since it does not use the normal assumption for random errors and more samples are needed to compensate for this.

When we increase the component variances σ_c^2 from 0.01 to 0.1, we found similar consistent results for the GWL estimator and biased results for the CS and SWL estimator. But when we increase σ_c^2, the bias of the CS estimator and the SWL estimator decreases. For example, with $n = 200$ the bias of the SWL estimator decreases from 0.1157 ($\sigma_c^2 = 0.01$) to 0.0684 with ($\sigma_c^2 = 0.1$). Note that the CS estimator and the SWL estimator are based on the normal distribution assumption for random errors. When σ_c^2 increases the bi-modal mixture of normal distributions tends to a uni-modal distribution which may be very close to some normal distribution. Thus the normal assumption for random errors may be reasonable for large values of σ_c^2 even if the distribution of ϵ is a mixture normal distribution.

Tsiatis and Davidian (2001) pointed out that the estimating equation of conditional score method may have multiple-roots. We also investigated the multiple roots problem for estimating equations of both methods. Typical score plots are shown in Figure 1. We can see that all methods have a solution close to the truth. In these typical score plots the generalized working likelihood estimating equations have a single root (the graph in the right panel of Figure 1). The SWL method has multiple solutions, which can be seen from the dotted line of the graph in the left panel of Figure 1 since the dotted line and the horizontal straight line have two intersection points. One close to -1 and the other one close to -20, which represents the two roots of the score function. The CS estimator may also have multiple solutions. When we solve the score functions, we use Newton-Raphson algorithms. From the dashed line of the graph in the left panel of Figure 1 we can see that if we choose the starting value of γ as 3, the Newton-Raphson algorithm will provide a sequence of γ-values converging to 20. In practice as Tsiatis and Davidian (2001) suggested we may choose the naive regression

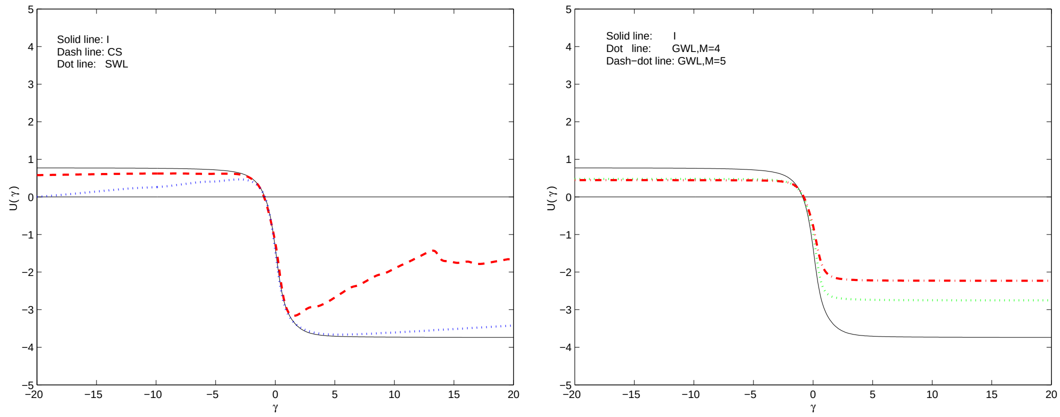

Figure 1: Typical score plots for simulation data sets. I, 'ideal' method; CS, conditional score method; SWL, simple working likelihood method; GWL, generalized working likelihood method.

estimate as starting value to locate the correct estimate.

We also find out that when using generalized working likelihood method and choosing $M = 4$, the score function may have an *outlier* solution. Note that all the outliers occurring in our simulations are smaller than -20 which is far away from the true estimates around -1.0 (outside more than 3 times of standard deviations). In Figure 2 the solution for $M = 4$ is an outlier while the solution for $M = 5$ is the truth. The outliers will result in poor estimate for standard error and means. Similarly as Song and Huang (2005), when all estimates exist and are non-outliers, general working likelihood estimates are stable and have small bias, regardless of the distributions of random effects and error terms.

For the GWL estimator, the value of M plays an important role in the accuracy of the estimators. To investigate this, we consider one scenario studied before where $(\omega_{i0}, \omega_{i1})$ and the error term are both mixtures of normal distributions. The maximum number of longitudinal observations is 24, thus the maximum value that M can take is 8. On the other hand, $\hat{\Psi}$ is based on $\xi_{i,r}(s)$ (or the estimate of $(\omega_{i0}, \omega_{i1})$) and σ^2, which need three observations for estimation. Therefore we consider different values for M ranging from 3 to 8. The results are summarized in Table 3. From the results we can see that when $M = 3$ the estimate has

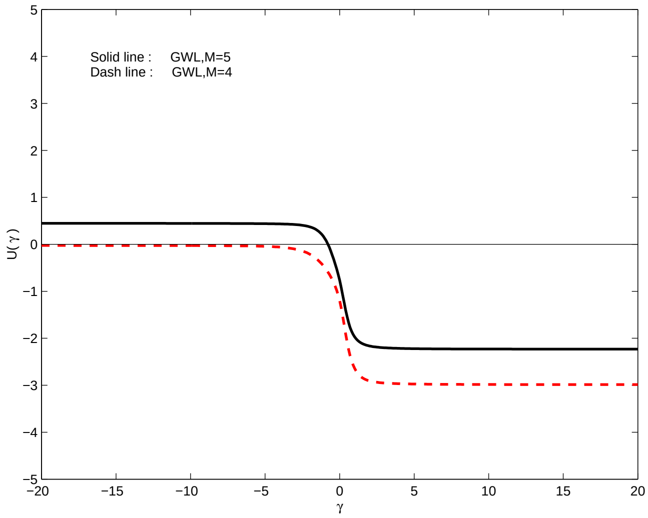

Figure 2: Typical score plots for a simulation data set. GWL, generalized working likelihood method.

Table 3: Comparisons for different values of M. The true parameter value is -1.0. Sample size in the table means the sample size used to estimate $\hat{\Psi}$. The total number of observations is 200.

	$M = 3$	$M = 4$	$M = 5$	$M = 6$	$M = 7$	$M = 8$
Estimate	-0.8823	-1.0274	-1.0172	-1.0066	-1.0031	-0.9811
SD	0.465	0.366	0.289	0.251	0.237	0.237
SE	0.567	0.423	0.288	0.255	0.244	0.233
Sample size	120	110	106	98	91	79

a larger bias comparing to the other estimates. This is because smaller M will give a large variance for the estimate of $\hat{\Psi}$ and then lead to a poor estimate of Ψ. When $M = 4$ the estimate becomes better although there is still a little bias. When $M = 5, 6, 7$ the estimate has very little bias. When M increases from 7 to 8, the number of subjects used to estimate $\hat{\Psi}$ decreases from 91 to 79 and the bias of the estimate increases slightly. Under this simulation scenario, more than 80 subjects used to estimate Ψ seems sufficient.

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. PBC is a chronic disease that can destroy some of the bile ducts linking your liver to your gut and then bile can no longer flow through them. Instead it builds up in the liver, damaging the liver cells and causing inflammation and

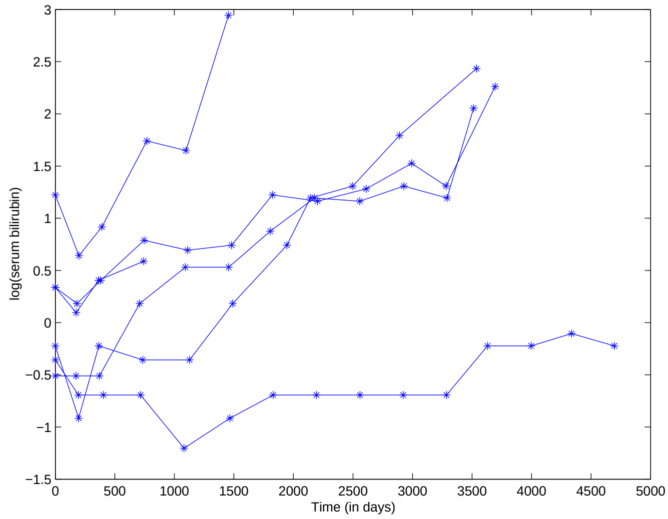

Figure 3: Longitudinal observation plot for six patients.

scarring. In the clinical trial, survival status and laboratory results such as serum bilirubin of 312 patients were recorded. In this clinical trial, 158 out of 312 patients took the drug D-penicillamine and the other patients are in the control group. Serum bilirubin is measured at irregularly-spaced time points, recorded until death or censoring. Among the 312 patients, the maximum number of repeated measurement is 16. More detailed discussion of the trial and data set can be found in Ding and Wang (2008) and Fleming and Harrington (1991).

We here consider the biomarker serum bilirubin as the latent process $W_i(s)$ and treatment type as the time-independent covariate Z_i in the Cox regression model $\lambda_i(s) = \lambda_0(s)\exp(\gamma W_i(s) + \beta Z_i)$ and longitudinal model $W_i(s) = \omega_{i0} + \omega_{i1}s$. In our analysis we took logarithm transformation on the serum bilirubin values as the longitudinal measurements. Six patients are randomly selected and their serum bilirubin values are plotted in Figure 3.

We fit the model using 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 randomly partitioned into

Table 4: 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)

three groups as replicated observations to estimate $\mathbf{\Psi}$. Thus we only used the 32 patients' longitudinal data, where each patient has more than 12 measurements, to estimate $\mathbf{\Psi}$. Note that although these 32 patients is just a small proportion of the total 312 patients, the descriptive statistics of their longitudinal measurements are very similar to the descriptive statistics for all patients' longitudinal measurements. For example the mean, standard deviation and range for these 32 patients' longitudinal data are 0.11, 0.40 and 2.10 respectively, while the corresponding values for all patients are 0.26, 0.48 and 2.61. After we obtain the estimate $\hat{\mathbf{\Psi}}$ we solve the score functions in (11) which are based on all observations.

The results are shown in Table 4. The three approaches provide similar results. All three methods suggest that the coefficient estimate $\hat{\gamma}$ for the latent process of serum bilirubin is non-significant. The coefficient estimate based on baseline serum bilirubin in Fleming and Harrington (1991) is 0.8, significantly different from 0. This suggests that the baseline serum bilirubin is a risk factor for survival times but 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 non-significant. From Table 4 we can see that all three methods give the same result that the coefficient estimate $\hat{\beta}$ for treatment Z_i is not significantly different from 0.

6 Discussion

We have proposed new methods for joint modelling of survival events and error-contaminated time-dependent processes. The estimators are easily computed and their large sample properties are shown. We suggest using the generalized working likelihood method in practice if the number of longitudinal measurements is relatively large for each subject, since it does not lose much efficiency when the random error is normally distributed and it results in the least bias if the error term is from a mixture of normal distributions.

When using the general working likelihood method, we need replicated observations. Even if replicated observations do not exist, we still obtain replicates from the longitudinal observations using the method in Section 4.4. When partitioning the $3M$ longitudinal measurements into three groups, we need to do it randomly for each individual. 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 doing in such way $(\widetilde{W}_{ij,r}, t_{ij,r}, j = 1, \cdots, M)$ will not have the same distribution for different values of r and the large sample properties of the estimator cannot be guaranteed.

Most of 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 methods in this paper are not valid. This is left as future research work.

A Regularity 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, $\sup_{s \in [0,\tau], \boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{E} |\boldsymbol{E}^{(k)}(W, \boldsymbol{\theta}, s) - \boldsymbol{e}^{(k)}(\boldsymbol{\theta}, s)|^{\otimes 2} \to 0$, as $n \to \infty$.

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

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

C.4 The matrix $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0, \tau) = \int \boldsymbol{v}(\boldsymbol{\theta}_0, s) e^{(0)}(\boldsymbol{\theta}_0, s)\lambda_0(s)ds$ is positive definite.

B Proof of Theorem 3.1

Proof. Under the regularity conditions in Appendix A, $l(\boldsymbol{\theta})$ 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 the regularity conditions in Chapter 8 of Fleming and Harrington (1991) $l(\boldsymbol{\theta})$ has a unique maximum at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ in a neighborhood of $\boldsymbol{\theta}_0$.

Now we prove the asymptotic normality of the estimator. The first-order Taylor extension for $\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, t)$ gives $\sqrt{n}\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, t) = \sqrt{n}\hat{\boldsymbol{\mathcal{I}}}(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2, \tau)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1)$, where $\hat{\boldsymbol{\mathcal{I}}}(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2, \tau) = \partial \hat{\boldsymbol{U}}^{(1)}/\partial\gamma\partial\boldsymbol{\beta} = -n^{-1}\sum_i \int_0^\tau \hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2, s)dN_i(s)$,

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

and

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

For any vector $\boldsymbol{a}$, the notation $\boldsymbol{a}^{\otimes 2}$ denotes the outer product $\boldsymbol{a}\boldsymbol{a}'$.

Under regularity conditions in Appendix A, $\hat{\boldsymbol{V}}(\boldsymbol{\theta}, \sigma^2, s)$ converges to $\boldsymbol{v}(\boldsymbol{\theta}, s)$ and $n^{-1}\int \hat{\boldsymbol{V}}(\boldsymbol{\theta}, \hat{\sigma}^2, s)\sum_i dN_i(s)$ converges to $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0, \tau) = \int \boldsymbol{v}(\boldsymbol{\theta}, s)e^{(0)}(\boldsymbol{\theta}, s)\lambda_0(s)ds$. If $\sqrt{n}\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, \tau)$ converges weakly to a normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_U)$, we have $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \Rightarrow N(\mathbf{0}, \boldsymbol{R}_1)$, where

$$\boldsymbol{R}_1 = \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0, \tau)^{-1}\boldsymbol{\Sigma}_U(\boldsymbol{\theta}_0, \sigma^2, \tau)\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0, \tau)^{-1}.$$

So we only need to show the asymptotic property for $\sqrt{n}\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, \tau)$.

Note that $\sum_i \int_0^t \left[\begin{pmatrix} \widehat{W}_i(s) - \hat{\sigma}^2 v_i(s)\gamma \\ \boldsymbol{Z}_i(s) \end{pmatrix} - \frac{\hat{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, s)}{\hat{E}^{(0)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, s)}\right]\hat{E}_i^{(0)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, s)\lambda_0(s)ds = \mathbf{0}$. We can write

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

If we define $\mathcal{F}_{i,s} = \sigma\{I[\tilde{T}_i \le u], \delta_i, \boldsymbol{\omega}_i, \boldsymbol{Z}_i(u), u \le s\}$ and $\mathcal{F}_s = \bigvee_i \mathcal{F}_{i,s}$, then we have $\mathcal{E}[dN_i(s) - \hat{E}_i^{(0)}(\boldsymbol{\theta}_0, \sigma^2, s)\lambda_0(s)ds|\mathcal{F}_s] = 0$, which implies that $dN_i(s) - \hat{E}_i^{(0)}(\boldsymbol{\theta}_0, \sigma^2, s)\lambda_0(s)ds$ is a martingale. Under regularity conditions, we also have that $\sup_{\boldsymbol{\theta}, s}\mathcal{E}\left|\frac{\boldsymbol{e}^{(1)}(\boldsymbol{\theta}, s)}{e^{(0)}(\boldsymbol{\theta}, s)} - \frac{\hat{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}, s)}{\hat{E}^{(0)}(\boldsymbol{\theta}, s)}\right|^{\otimes 2} = o(1)$. This implies that $III = o_p(1)$. 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 can show that $\sqrt{n}\hat{\boldsymbol{U}}^{(1)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, t)$ converges weakly to a normal distribution with mean $\mathbf{0}$ and covariance matrix

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

□

Define

$$\hat{\boldsymbol{\Phi}}_i(\boldsymbol{\theta}_0, \hat{\sigma}^2, t) = \int_0^t \left[\begin{pmatrix} \widehat{W}_i(s) \\ \boldsymbol{Z}_i(s) \end{pmatrix} - \frac{\hat{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, s)}{\hat{E}^{(0)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, s)} \right] dN_i(s) - \int_0^t \left[\begin{pmatrix} \widehat{W}_i(s) - \hat{\sigma}^2 v_i(s)\gamma \\ \boldsymbol{Z}_i(s) \end{pmatrix} - \frac{\hat{\boldsymbol{E}}^{(1)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, s)}{\hat{E}^{(0)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, s)} \right] \hat{E}_i^{(0)}(\boldsymbol{\theta}_0, \hat{\sigma}^2, s)\hat{\lambda}_0(s)ds. \tag{12}$$

we have that an estimator for $\boldsymbol{\Sigma}_U(\boldsymbol{\theta}_0, \sigma^2, t)$ is

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

Thus an estimator for $\boldsymbol{R}_1$ is $\hat{\boldsymbol{R}}_1 = \hat{\boldsymbol{\mathcal{I}}}(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2, \tau)^{-1}\hat{\boldsymbol{\Sigma}}_U(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2, \tau)\hat{\boldsymbol{\mathcal{I}}}(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2, \tau)^{-1}$.

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

D Asymptotic distribution for the GWL estimator

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

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

and

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

We can show that $\sqrt{n}(\breve{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges weakly to a normal distribution $N(0, \boldsymbol{R}_2)$. Let $\breve{\lambda}_0(s)ds = d\bar{N}(s)/\breve{E}^{(0)}(\breve{\boldsymbol{\theta}}, s)$ and

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

An estimate for $\boldsymbol{R}_2$ is

$$\breve{\boldsymbol{R}}_2 = \breve{\mathcal{I}}(\breve{\boldsymbol{\theta}}, \hat{\boldsymbol{\Psi}}, \tau)^{-1} \left[\frac{1}{n} \sum_i \breve{\boldsymbol{\Phi}}_i(\breve{\boldsymbol{\theta}}, \hat{\boldsymbol{\Psi}}, \tau)^{\otimes 2} + 2 \sum_{i=2}^n \breve{\boldsymbol{\Phi}}_i(\breve{\boldsymbol{\theta}}, \hat{\boldsymbol{\Psi}}, \tau) \otimes \breve{\boldsymbol{\Phi}}_1(\breve{\boldsymbol{\theta}}, \hat{\boldsymbol{\Psi}}, \tau) \right] \breve{\mathcal{I}}(\breve{\boldsymbol{\theta}}, \hat{\boldsymbol{\Psi}}, \tau)^{-1}.$$

References

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

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

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

Faucett C. L. and Thomas D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Statstics in Medicine*, **15**:1663-1685.

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.

Prentice R. (1982). Covariate measurement errors and parameter estimation in a failure time regression models. *Biometrika*, **69**:331-342.

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

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

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

Tsiatis 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. *Journal of the American Statistical Association*, **90**:27-37.

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 longitudinal data measured with error. *Biometrics*, **53**:330-339.