Statistics and Its Interface Volume 9 (2016) 333–341

Semiparametric random effects models for longitudinal data with informative observation times

Yang Li* and Yanqing Sun

Longitudinal data frequently arise in many fields such as medical follow-up studies focusing on specific longitudinal responses. In such situations, the responses are recorded only at discrete observation times. Most existing approaches for longitudinal data analysis assume that the observation or follow-up times are independent of the underlying response process, either completely or given some known covariates. We present a joint analysis approach in which possible correlations among the responses, observation and follow-up times can be characterized by time-dependent random effects. Estimating equations are developed for parameter estimation and the resulting estimates are shown to be consistent and asymptotically normal. A simulation study is conducted to assess the finite sample performance of the approach and the method is applied to data arising from a skin cancer study.

Keywords and phrases: Estimating equations, Informative censoring, Informative observation process, Joint analysis approach, Longitudinal data.

1. INTRODUCTION

Longitudinal data arise in many fields such as medical follow-up studies that focus on longitudinal responses. In such situations, each study subject is observed only at finite discrete times rather than continuously. Therefore, the responses are known only at a set of observation times but missing otherwise. The resulting data are usually incomplete and unbalanced among individuals.

Analysis of longitudinal data concerns two processes: one is the underlying response process, which is usually of practical interest but not continuously observable. The other refers to the observation process, which determines the discrete observation times. Many authors have considered the analysis of longitudinal data, for example, Diggle et al. (1994) who presented a relatively comprehensive review about the commonly considered models and estimation methods. Lin and Ying (2001), Welsh et al. (2002), Wellner and Zhang (2007) and Sun (2010) developed some semiparametric and nonparametric procedures for regression analysis. These approaches all assume that the two processes mentioned above are independent, either completely or conditional on some known covariates. To relax this assumption, Sun et al. (2007b), Zhao and Tong (2011) and Zhao et al. (2013b) modeled the possible correlations by time-independent random effects. However, these methods assume follow-up times to be independent from both the response and the observation processes given covariates.

In many situations, the underlying response process, the observation and follow-up times may be correlated. For example, both observation times and responses may depend on the stage of disease progression, which can also often determine the follow-up time. Lipsitz et al. (2002) considered general linear models for longitudinal data where the responses were assumed to have a multivariate Gaussian distribution. Sun et al. (2007), He et al. (2009) and Sun et al. (2012) proposed joint model based approaches; however, it is assumed that the shared random effects are fixed over time or follow some specific distributions, and the covariates are either multiplicative or additive in their effects to the response process. Without such specific distribution assumption, Sun et al. (2005) and Zhao et al. (2013) considered marginal model based methods; however, the models indicate that when the observation process is common for everyone, people with the same covariates are expected to have the same responses throughout the study. It is apparent that such assumptions may not be realistic in many applications.

We present a joint analysis approach for longitudinal data by which the possible correlations can be characterized by time-dependent random effects with arbitrary distributions. For the response process, a class of semiparametric transformation models are considered. Estimating equations are developed for parameter estimation and the resulting estimators are shown to be consistent and asymptotically normal. The remainder of this paper is organized as follows. We introduce notation and present the relevant models in Section 2. Section 3 presents the estimation procedure and establishes asymptotic properties of the proposed estimators. In Section 4, we demonstrate a model-checking technique and an extensive simulation study is presented in Section 5 to evaluate finite sample properties of the estimation procedure. An illustrative example is given in Section 6 and some discussion and remarks are provided in Section 7.

*Corresponding author.

2. NOTATION AND MODELS

Consider a longitudinal study in which subjects are observed only at discrete times. For subject i $(i = 1, \ldots, n)$, let $Y_i(t)$ denote the response process and let $N_i(t)$ be the observation process which gives the cumulative number of observations at time t. In practice, one often observes $\widetilde{N}_i(t) = N_i(t \wedge C_i)$ where $a \wedge b = min(a, b)$ and C_i denotes a possible censoring or follow-up time. Let $\{T_{i,1}, \cdots, T_{i,m_i}\}$ be the discrete times when $Y_i(t)$ is observed and let $\mathbf{Z}_i(t)$ be a p-dimensional vector of covariates, assumed to be continuously traceable in the study. In the following, we present a joint modeling approach and model the possible correlation between $Y_i(t)$, $N_i(t)$ and C_i through an unobserved random process $\mathbf{b}_i(t) = (b_{1i}(t), b_{2i}(t), b_{3i}(t))'$. Define $\mathcal{B}_{it} = \{\mathbf{b}_i(s), s \le t\}$ and $\mathcal{Z}_{it} = \{\mathbf{Z}_i(s), s \le t\}$. We assume that the $\mathbf{b}_i(t)$'s are independent and identically distributed with $b_{1i}(t) > 0$ and $b_{2i}(t) > 0$, $\mathcal{B}_{it}$ is independent of $\mathcal{Z}_{it}$, and given $\mathcal{Z}_{it}$ and $\mathcal{B}_{it}$, C_i, $N_i(t)$ and $Y_i(t)$ are mutually independent. Also we assume that the mean function of $Y_i(t)$ can be postulated by the following semiparametric transformation model

$$E\{Y_i(t)|\mathbf{Z}_i(t), \mathbf{b}_i(t)\} = g\{\mu_0(t)e^{\theta'\mathbf{Z}_i(t)}\}b_{1i}(t), \tag{1}$$

where $g(\cdot)$ is a known twice continuously differentiable and strictly increasing link function, θ is a vector of unknown regression parameters and $\mu_0(t)$ denotes an unspecified smooth function of t. We assume that $E\{b_{1i}(t)|d\widetilde{N}_i(t) = 1, \mathcal{Z}_{it}\} = 1$ for identifiability. In particular, when $g(x) = x$, $\mu_0(t)$ represents the baseline mean function that is estimable at $\{T_{i,1}, \cdots, T_{i,m_i}\}$.

The observation process $N_i(t)$ is assumed to follow the marginal proportional rates model given by

$$E\{dN_i(t)|\mathbf{Z}_i(t), \mathbf{b}_i(t)\} = \exp\{\gamma'\mathbf{Z}_i(t)\}b_{2i}(t)d\Lambda_0(t), \tag{2}$$

where $E\{b_{2i}(t)\} = 1$, γ is a vector of unknown parameters and $d\Lambda_0(t)$ is an unknown baseline rate function. It can be seen that both of the above models can be viewed as natural generalizations of the transformation model and proportional rates model studied in Li et al. (2010), Zhao et al. (2011) and Zhao et al. (2013) among others. Compared with the existing models, the proposed models are relatively flexible in handling the possible dependence since neither the form nor the distribution of $\mathbf{b}_i(t)$ needs to be specified. By taking different forms of $g(\cdot)$ and $\mathbf{b}_i(t)$, model (1) allows for various types of dependence for the mean function of $Y_i(t)$ on $N_i(t)$ and $\mathbf{Z}_i(t)$. In particular, when either $b_{1i}(t)$ or $b_{2i}(t)$ is unity or independent of the other one, the two processes $Y_i(t)$ and $N_i(t)$ are independent given $\mathcal{Z}_{it}$. Therefore, the estimation procedure proposed next also applies to data with noninformative observation times as special cases.

For the follow-up or censoring time C_i, we consider the following additive hazards model

$$\lambda_i(t|\mathbf{Z}_i(t), \mathbf{b}_i(t)) = \lambda_0(t) + \xi'\mathbf{Z}_i(t) + b_{3i}(t), \tag{3}$$

where $E\{b_{3i}(t)\} = 0$, $\lambda_0(t)$ is an unknown baseline hazard function and ξ is an unknown vector of regression parameters. The random effects $b_{1i}(t)$, $b_{2i}(t)$ and $b_{3i}(t)$ characterize possible correlations between C_i and $Y_i(t)$, $N_i(t)$, for which $b_{3i}(t) = 0$ implies noninformative censoring. The same model has also been studied in Kalbfleisch and Prentice (2002), Lin et al. (1998), Zhang et al. (2005) and Sun et al. (2013) among others. In the following, we study the joint analysis of the proposed models with the focus on estimation of regression parameters θ along with γ and ξ.

3. ESTIMATION PROCEDURE

In this section, we present an estimation procedure for θ which is usually of primary interest. To this end, first note that $\widetilde{N}_i(t)$ jumps by one at time t if and only if $C_i \ge t$ and $dN_i(t) = 1$. Based on the conditional independence assumption between C_i, $N_i(t)$ and $Y_i(t)$ given $\mathcal{Z}_{it}$ and $\mathcal{B}_{it}$, we have, under (2)

$$\begin{aligned} E\{d\widetilde{N}_i(t)|\mathcal{Z}_{it}\} &= E\left[E\{I(t \le C_i)dN_i(t)|\mathcal{Z}_{it}, \mathcal{B}_{it}\}\middle|\mathcal{Z}_{it}\right] \\ &= E\left[E\{I(t \le C_i)|\mathcal{Z}_{it}, \mathcal{B}_{it}\}E\{dN_i(t)|\mathcal{Z}_{it}, \mathcal{B}_{it}\}\middle|\mathcal{Z}_{it}\right] \\ &= E\{I(t \le C_i)b_{2i}(t)|\mathcal{Z}_{it}\}\exp\{\gamma'\mathbf{Z}_i(t)\}d\Lambda_0(t). \end{aligned} \tag{4}$$

By the property of double expectation and model (3), the first term in (4) equals

$$\begin{aligned} & E\{I(t \le C_i)b_{2i}(t)|\mathcal{Z}_{it}\} \\ = \; & E\left\{exp\{-\Lambda_0^*(t) - B_i(t) - \xi'\mathbf{Z}_\mathbf{i}^*(t)\}b_{2i}(t)\middle|\mathcal{Z}_{it}\right\}, \end{aligned}$$

where $\Lambda_0^*(t) = \int_0^t \lambda_0(s)ds$, $B_i(t) = \int_0^t b_{3i}(s)ds$ and $Z_i^*(t) = \int_0^t \mathbf{Z}_i(s)ds$. Hence,

$$E\{d\widetilde{N}_i(t)|\mathcal{Z}_{it}\} = \exp\{\eta'\mathbf{X}_\mathbf{i}^*(\mathbf{t})\}d\Lambda_1^*(t), \tag{5}$$

where $\eta = (\gamma', \xi')'$, $\mathbf{X}_i^*(t) = (\mathbf{Z}_\mathbf{i}'(t), -\mathbf{Z}_\mathbf{i}'^*(t))'$ and $d\Lambda_1^*(t) = \exp\{-\Lambda_0^*(t)\}E[b_{2i}(t)exp\{-B_i(t)\}]d\Lambda_0(t)$.

Let τ be a known constant representing the length of the study. Define $dM_i^*(t;\eta) = d\widetilde{N}_i(t) - e^{\eta'\mathbf{X}_\mathbf{i}^*(t)}d\Lambda_1^*(t)$ and $dM_i^*(t) = dM_i^*(t;\eta_0)$, where η_0 denotes the true value of η. It is straightforward to show that $M_i^*(t)$ is a mean-zero stochastic process. It follows that η and $\Lambda_1^*(t)$ can be estimated by $\hat{\eta}$ and $\hat{\Lambda}_1^*(t;\hat{\eta})$, respectively, by solving the following two estimating equations

$$U_\eta(\eta) = \sum_{i=1}^n \int_0^\tau \left\{\mathbf{X}_i^*(t) - \bar{X}^*(t;\eta)\right\}d\widetilde{N}_i(t) = 0, \tag{6}$$

and

$$\sum_{i=1}^n \left[d\widetilde{N}_i(t) - e^{\eta'\mathbf{X}_i^*(t)}d\Lambda_1^*(t)\right] = 0, \tag{7}$$

where $\bar{X}^*(t;\eta) = S^{(1)}(t;\eta)/S^{(0)}(t;\eta)$ and $S^{(k)}(t;\eta) = n^{-1}\sum_{i=1}^n e^{\eta' \mathbf{X}_i^*(t)} \mathbf{X}_i^*(t)^{\otimes k}$ for $k = 0, 1$ and 2. Here and throughout $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$ and $a^{\otimes 2} = aa'$. Define $\hat{\Lambda}_1^*(t) = \hat{\Lambda}_1^*(t;\hat{\eta})$, $\bar{x}^*(t) = lim_{n\to\infty}\bar{X}^*(t;\eta_0)$ and $s^{(k)}(t) = lim_{n\to\infty}S^{(k)}(t;\eta_0)$.

For the estimation of θ, consider

$$\begin{aligned} & E\{Y_i(t)d\widetilde{N}_i(t)|\mathcal{Z}_{it}\} \\ = \; & E\{Y_i(t)I\big(d\widetilde{N}_i(t)=1\big)|\mathcal{Z}_{it}\} \\ = \; & \frac{E\{Y_i(t)I\big(d\widetilde{N}_i(t)=1\big)|\mathcal{Z}_{it}\}}{E\{I\big(d\widetilde{N}_i(t)=1\big)|\mathcal{Z}_{it}\}} E\{d\widetilde{N}_i(t)|\mathcal{Z}_{it}\} \end{aligned}$$

by the definition of $d\widetilde{N}_i(t)$ and simple manipulation. From the conditional independence assumption between C_i, $N_i(t)$ and $Y_i(t)$ given $\mathcal{Z}_{it}$ and $\mathcal{B}_{it}$, the last equality equals

$$\begin{aligned} & E\{Y_i(t)d\widetilde{N}_i(t)|\mathcal{Z}_{it}\} \\ = \; & \frac{E\big[E\{Y_i(t)I\big(d\widetilde{N}_i(t)=1\big)|\mathcal{Z}_{it},\mathcal{B}_{it}\}|\mathcal{Z}_{it}\big]}{E\{I\big(d\widetilde{N}_i(t)=1\big)|\mathcal{Z}_{it}\}} E\{d\widetilde{N}_i(t)|\mathcal{Z}_{it}\} \\ = \; & \frac{g\{\mu_0(t)e^{\theta'\mathbf{Z}_i(t)}\}E\{b_{1i}(t)I\big(d\widetilde{N}_i(t)=1\big)|\mathcal{Z}_{it}\}}{E\{I\big(d\widetilde{N}_i(t)=1\big)|\mathcal{Z}_{it}\}} \\ & \times E\{d\widetilde{N}_i(t)|\mathcal{Z}_{it}\} \\ = \; & g\{\mu_0(t)e^{\theta'\mathbf{Z}_i(t)}\}E\{b_{1i}(t)|d\widetilde{N}_i(t)=1,\mathcal{Z}_{it}\} \\ & \times E\{d\widetilde{N}_i(t)|\mathcal{Z}_{it}\} \\ = \; & g\{\mu_0(t)e^{\theta'\mathbf{Z}_i(t)}\}E\{d\widetilde{N}_i(t)|\mathcal{Z}_{it}\}, \end{aligned}$$

under model (1). Combining (5), it follows that

$$(8)\quad E\{Y_i(t)d\widetilde{N}_i(t)|\mathcal{Z}_{it}\} = e^{\eta'\mathbf{X}_i^*(t)}g\{\mu_0(t)e^{\theta'\mathbf{Z}_i(t)}\}\,d\Lambda_1^*(t).$$

We define

$$dM_i(t;\theta,\eta) = Y_i(t)d\widetilde{N}_i(t) - e^{\eta'\mathbf{X}_i^*(t)}g\{\mu_0(t)e^{\theta'\mathbf{Z}_i(t)}\}\,d\Lambda_1^*(t)$$

and $dM_i(t) = dM_i(t;\theta_0,\eta_0)$, where θ_0 denotes the true value of θ. Then $M_i(t)$ is a mean-zero stochastic process, which naturally suggests the following estimating equations to estimate θ and $\mu_0(t)$:

$$\begin{aligned} (9)\qquad & \sum_{i=1}^n \left[Y_i(t)d\widetilde{N}_i(t) - e^{\hat{\eta}'\mathbf{X}_i^*(t)}g\{\mu_0(t)e^{\theta'\mathbf{Z}_i(t)}\}d\hat{\Lambda}_1^*(t)\right] \\ = \; & 0,\ 0 \le t \le \tau, \end{aligned}$$

and

$$\begin{aligned} (10)\qquad & \sum_{i=1}^n \int_0^\tau W(t)\mathbf{Z}_i(t) \\ \times \; & \left[Y_i(t)d\widetilde{N}_i(t) - e^{\hat{\eta}'\mathbf{X}_i^*(t)}g\{\mu_0(t)e^{\theta'\mathbf{Z}_i(t)}\}d\hat{\Lambda}_1^*(t)\right] = 0, \end{aligned}$$

where $W(t)$ is a possibly data-dependent weight function. We denote the estimates of θ and $\mu_0(t)$ by $\hat{\theta}$ and $\hat{\mu}_0(t;\hat{\theta},\hat{\eta})$, respectively. Define $\hat{\mu}_0(t) = \hat{\mu}_0(t;\hat{\theta},\hat{\eta})$.

In general, neither $\hat{\theta}$ nor $\hat{\mu}_0(t)$ have closed forms and some iterative algorithms may be necessary to solve (9) and (10). For some special cases, $\hat{\mu}_0(t)$ can be written explicitly. For example, when $g(x) = \log(x)$, it can be shown that

$$\hat{\mu}_0(t) = \exp\left\{\frac{\sum_{i=1}^n Y_i(t)d\widetilde{N}_i(t)}{\sum_{i=1}^n e^{\hat{\eta}'\mathbf{X}_i^*(t)}d\hat{\Lambda}_1^*(t)} - \hat{\theta}'\bar{Z}(t;\hat{\eta})\right\}$$

and

$$\begin{aligned} \hat{\theta} = & \left\{\sum_{i=1}^n \int_0^\tau W(t)\{\mathbf{Z}_i(t) - \bar{Z}(t;\hat{\eta})\}\mathbf{Z}_\mathbf{i}'(t)e^{\hat{\eta}'\mathbf{X}_i^*(t)}d\hat{\Lambda}_1^*(t)\right\}^{-1} \\ & \times \sum_{i=1}^n \int_0^\tau W(t)\{\mathbf{Z}_i(t) - \bar{Z}(t;\hat{\eta})\}Y_i(t)d\widetilde{N}_i(t), \end{aligned}$$

where $\bar{Z}(t;\hat{\eta}) = \frac{\sum_{i=1}^n \mathbf{Z}_i(t)e^{\hat{\eta}'\mathbf{X}_i^*(t)}}{\sum_{i=1}^n e^{\hat{\eta}'\mathbf{X}_i^*(t)}}$.

To establish the asymptotic properties of $\hat{\theta}$, we define

$$\widehat{M}_i^*(t) = \widetilde{N}_i(t) - \int_0^t e^{\hat{\eta}'\mathbf{X}_i^*(s)}d\hat{\Lambda}_1^*(s),$$

$$\widehat{M}_i(t) = \int_0^t Y_i(s)d\widetilde{N}_i(s) - \int_0^t e^{\hat{\eta}'\mathbf{X}_i^*(s)}g\{\hat{\mu}_0(s)e^{\hat{\theta}'\mathbf{Z}_i(s)}\}d\hat{\Lambda}_1^*(s),$$

$$\hat{E}_Z(t;\hat{\theta},\hat{\eta}) = \frac{\sum_{i=1}^n \mathbf{Z}_i(t)\dot{g}\{\hat{\mu}_0(t)e^{\hat{\theta}'\mathbf{Z}_i(t)}\}e^{\hat{\theta}'\mathbf{Z}_i(t)+\hat{\eta}'\mathbf{X}_i^*(t)}}{\sum_{i=1}^n \dot{g}\{\hat{\mu}_0(t)e^{\hat{\theta}'\mathbf{Z}_i(t)}\}e^{\hat{\theta}'\mathbf{Z}_i(t)+\hat{\eta}'\mathbf{X}_i^*(t)}}$$

$e_z(t) = lim_{n\to\infty}\hat{E}_Z(t;\theta_0,\eta_0)$ and $\hat{E}_Z(t) = \hat{E}_Z(t;\hat{\theta},\hat{\eta})$.

The following theorem establishes the consistency and asymptotic normality of $\hat{\theta}$ and $\hat{\eta}$.

Theorem 1. *Assume that the conditions (C1)–(C5) given in the Appendix hold. Then $\hat{\theta}$ and $\hat{\eta}$ are consistent estimators of θ_0 and η_0, respectively. $n^{1/2}(\hat{\theta}-\theta_0)$ and $n^{1/2}(\hat{\eta}-\eta_0)$ converge weakly to mean-zero normal distributions with covariance matrices that can be consistently estimated by $\widehat{\Sigma}_\theta = \hat{A}_\theta^{-1}\widehat{\Sigma}\hat{A}_\theta^{-1}$ and $\widehat{\Sigma}_\eta = \hat{\Omega}_\eta^{-1}\widehat{\Psi}\hat{\Omega}_\eta^{-1}$, respectively, where $\widehat{\Sigma} = n^{-1}\sum_{i=1}^n(\hat{\xi}_{1i} - \hat{\xi}_{2i} - \hat{\xi}_{3i})^{\otimes 2}$, $\widehat{\Psi} = n^{-1}\sum_{i=1}^n \hat{\zeta}_i^{\otimes 2}$,*

$$\begin{aligned} \hat{\xi}_{1i} &= \int_0^\tau W(t)\Big(\mathbf{Z}_i(t) - \hat{E}_Z(t)\Big)d\widehat{M}_i(t)\,, \\ \hat{\xi}_{2i} &= \int_0^\tau \frac{W(t)\hat{D}(t;\hat{\theta},\hat{\eta})}{S^{(0)}(t;\hat{\eta})}d\widehat{M}_i^*(t)\,, \\ \hat{\xi}_{3i} &= \int_0^\tau \hat{A}_\eta\hat{\Omega}_\eta^{-1}\Big(\mathbf{X}_i^*(t) - \bar{X}^*(t;\hat{\eta})\Big)d\widehat{M}_i^*(t)\,, \\ \hat{\zeta}_i &= \int_0^\tau \Big(\mathbf{X}_i^*(t) - \bar{X}^*(t;\hat{\eta})\Big)d\widehat{M}_i^*(t)\,, \\ \hat{A}_\theta &= \frac{1}{n}\sum_{i=1}^n \int_0^\tau W(t)\dot{g}\{\hat{\mu}_0(t)e^{\hat{\theta}'\mathbf{Z}_i(t)}\} \\ \times \; & \{\mathbf{Z}_i(t) - \hat{E}_Z(t)\}^{\otimes 2}e^{\hat{\theta}'\mathbf{Z}_i(t)+\hat{\eta}'\mathbf{X}_i^*(t)}\hat{\mu}_0(t)d\hat{\Lambda}_1^*(t), \\ \hat{D}(t;\hat{\theta},\hat{\eta}) &= \frac{1}{n}\sum_{i=1}^n\{\mathbf{Z}_i(t) - \hat{E}_Z(t)\}g\{\hat{\mu}_0(t)e^{\hat{\theta}'\mathbf{Z}_i(t)}\}e^{\hat{\eta}'\mathbf{X}_i^*(t)}, \end{aligned}$$

$$\hat{A}_\eta = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau W(t) g\{\hat{\mu}_0(t)e^{\hat{\theta}'\mathbf{Z}_i(t)}\}e^{\hat{\eta}'\mathbf{X}_i^*(t)}$$
$$\times \quad \{\mathbf{Z}_i(t) - \hat{E}_Z(t)\}\{\mathbf{X}_i^*(t) - \bar{X}^*(t;\hat{\eta})\}' d\hat{\Lambda}_1^*(t),$$
$$\hat{\Omega}_\eta = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \{\mathbf{X}_i^*(t) - \bar{X}^*(t;\hat{\eta})\}^{\otimes 2} e^{\hat{\eta}'\mathbf{X}_i^*(t)} d\hat{\Lambda}_1^*(t).$$

The proof of the theorem above is sketched in Appendix A.

4. MODEL CHECKING

As mentioned above, a main advantage of the proposed methodology is that it is applicable to a class of correlated models through the link function $g(\cdot)$ and random effects $\mathbf{b}_i(t)$. On the other hand, one may question how to choose an appropriate form of $g(\cdot)$ for the response process. To answer this question, one may develop some model selection procedure and choose an optimal $g(\cdot)$ among several candidate models. However, such a strategy can be very difficult for longitudinal data because of their incompleteness. To access the adequacy of the proposed models with a given link function $g(\cdot)$, one can develop an omnibus goodness-of-fit test based on the cumulative summation of the residual process (Lin et al., 1993; Lin et al., 2000; Li et al., 2010; Zhao et al., 2013) as follows

$$\mathcal{F}(t,x) = n^{-1/2}\sum_{i=1}^{n}\int_0^t I(\mathbf{Z}_i(s) \le z) d\widehat{M}_i(s),$$

where $\{\mathbf{Z}_i(u) \le z\}$ represents that each component of $Z_i(u)$ is no greater than the corresponding component of z. In general, the distribution of $\mathcal{F}(t,x)$ is unknown or very difficult to obtain. Under the proposed models, $\mathcal{F}(t,x)$ is expected to flunctuate randomly around 0. In Appendix B, it is shown that the null distribution of $\mathcal{F}(t,x)$ can be approximated by a mean-zero Gaussian distribution

(11)
$$\widehat{\mathcal{F}}(t,z) = n^{-1/2}\sum_{i=1}^{n}\Big\{\hat{u}_{1i}(t,z) - \hat{u}_{2i}(t,z) - \hat{V}_\eta(t,z)\hat{\Omega}_\eta^{-1}\hat{\zeta}_i$$
$$- \hat{V}_\theta(t,z)\hat{A}_\theta^{-1}(\hat{\xi}_{1i} - \hat{\xi}_{2i} - \hat{\xi}_{3i})\Big\}e_i,$$

where $e_1, e_2, \ldots, e_n$ are independent standard normal variables independent of the observed data,

$$\hat{u}_{1i}(t,z) = \int_0^t \Big\{I\big(\mathbf{Z}_i(s) \le z\big) - \hat{E}_I(s,z;\hat{\theta},\hat{\eta})\Big\} d\widehat{M}_i(s),$$
$$\hat{u}_{2i}(t,z) = \int_0^t \frac{\hat{\Gamma}(s;\hat{\theta},\hat{\eta})}{S^{(0)}(s;\hat{\eta})} d\widehat{M}_i^*(s),$$
$$\hat{\Gamma}(t;\hat{\theta},\hat{\eta}) = n^{-1}\sum_{i=1}^{n}\{I(\mathbf{Z}_i(t) \le z) - \hat{E}_I(t,z;\hat{\theta},\hat{\eta})\}$$
$$\times\, g\{\hat{\mu}_0(t)e^{\hat{\theta}'\mathbf{Z}_i(t)}\}e^{\hat{\eta}'\mathbf{X}_i^*(t)},$$
$$\hat{V}_\eta(t,z)$$
$$= n^{-1}\sum_{i=1}^{n}\int_0^t g\{\hat{\mu}_0(s)e^{\hat{\theta}'\mathbf{Z}_i(s)}\}e^{\hat{\eta}'\mathbf{X}_i^*(s)}\{I(\mathbf{Z}_i(s) \le z)$$
$$- \hat{E}_I(s,z;\hat{\theta},\hat{\eta})\} \times \{\mathbf{X}_i^*(s) - \bar{X}^*(s;\hat{\eta})\}' d\hat{\Lambda}_1^*(s),$$
$$\hat{V}_\theta(t,z) = n^{-1}\sum_{i=1}^{n}\int_0^t \dot{g}\{\hat{\mu}_0(s)e^{\hat{\theta}'\mathbf{Z}_i(s)}\}I(\mathbf{Z}_i(s) \le z)$$
$$\times \{\mathbf{Z}_i(s) - \hat{E}_Z(s)\}' e^{\hat{\theta}'\mathbf{Z}_i(s) + \hat{\eta}'\mathbf{X}_i^*(s)}\hat{\mu}_0(s) d\hat{\Lambda}_1^*(s),$$
$$\hat{E}_I(t,z;\hat{\theta},\hat{\eta})$$
$$= \frac{\sum_{i=1}^{n} I(\mathbf{Z}_i(t) \le z)\dot{g}\{\hat{\mu}_0(t)e^{\hat{\theta}'\mathbf{Z}_i(t)}\}e^{\hat{\theta}'\mathbf{Z}_i(t) + \hat{\eta}'\mathbf{X}_i^*(t)}}{\sum_{i=1}^{n}\dot{g}\{\hat{\mu}_0(t)e^{\hat{\theta}'\mathbf{Z}_i(t)}\}e^{\hat{\theta}'\mathbf{Z}_i(t) + \hat{\eta}'\mathbf{X}_i^*(t)}}$$
$$e_I(t,z) = lim_{n\to\infty} E_I(t,z;\theta_0,\eta_0)$$

and $\hat{\zeta}_i$, $\hat{\xi}_{1i}$, $\hat{\xi}_{2i}$, $\hat{\xi}_{3i}$ are the same as defined in the previous section. Therefore for a given set of data, one can obtain a large number of realizations from $\widehat{\mathcal{F}}(t,z)$ by repeatedly generating standard normal random samples $\{e_1, e_2, \ldots, e_n\}$. A formal goodness-of-fit test can be performed with the corresponding p-value being calculated by comparing $sup_{0\le t\le\tau,z}|\mathcal{F}(t,z)|$ to a large number of realizations from $sup_{0\le t\le\tau,z}|\widehat{\mathcal{F}}(t,z)|$.

5. A SIMULATION STUDY

In this section, we present results obtained from an extensive simulation study conducted to assess the finite sample behavior of the estimation procedure proposed in the previous sections. In the study, the covariate Z_i was assumed to be a Bernoulli random variable with the probability of success being 0.5. Given Z_i and some unobserved random effects $\mathbf{b}_i(t) = (b_{1i}(t), b_{2i}(t), b_{3i}(t))'$, the hazard function of the censoring time C_i was assumed to have the form

(12)
$$\lambda_i(t|Z_i, \mathbf{b}_i(t)) = \lambda_0 - \xi Z_i + b_{3i}(t),$$

with the length of study τ being 1. The number of observations $N_i(t)$ was assumed to follow a Poisson process on $(0, C_i)$ with the rate function

(13)
$$E\{dN_i(t)|Z_i, \mathbf{b}_i(t)\} = \exp\{\gamma Z_i\} b_{2i}(t) d\Lambda_0(t)\,.$$

In practice, the exact time of C_i may not be observable and $d\widetilde{N}_i(t)$ is observed instead of $dN_i(t)$; thus we considered $E\{\widetilde{N}_i(t)|Z_i, \mathcal{B}_{it}\}$ for the observation times. From (12) and (13),

$$E\{d\widetilde{N}_i(t)|Z_i, \mathcal{B}_{it}\} = \exp\{\gamma Z_i + \xi Z_i t\} d\Lambda_1^*(t),$$

where $d\Lambda_1^*(t) = \exp\{-\lambda_0 t - B_i(t)\} b_{2i}(t) d\Lambda_0(t)$. Given Z_i and $\mathbf{b}_i(t)$, $\widetilde{N}_i(t)$ was assumed to follow a nonhomogeneous Poisson process and the total number of observation times m_i was generated with mean $E\{m_i\} = E\{\widetilde{N}_i(\tau)|Z_i, \mathcal{B}_{i\tau}\}$. Then the observation times $\{T_{i,1}, \ldots, T_{i,m_i}\}$ were taken as m_i order statistics from the density function

$$f_{\widetilde{N}}(t) = \frac{\exp\{\gamma Z_i + \xi Z_i t\} d\Lambda_1^*(t)}{\int_0^\tau \exp\{\gamma Z_i + \xi Z_i t\} d\Lambda_1^*(t)}.$$

Table 1. Estimation results for θ with the link function $g(x) = \log(x)$

	$n = 100$			$n = 200$		
θ_0	0	0.2	0.5	0	0.2	0.5
			$(\gamma_0, \xi_0) = (0, 0)$			
Bias	-0.004	-0.006	-0.018	-0.003	0.006	-0.008
SEE	0.186	0.199	0.218	0.131	0.140	0.154
SSE	0.193	0.208	0.220	0.129	0.149	0.149
CP	0.944	0.935	0.943	0.958	0.943	0.962
			$(\gamma_0, \xi_0) = (0, 0.2)$			
Bias	0.033	0.029	0.024	0.021	0.028	0.024
SEE	0.180	0.192	0.211	0.129	0.137	0.152
SSE	0.187	0.206	0.214	0.133	0.137	0.155
CP	0.939	0.929	0.953	0.942	0.947	0.948
			$(\gamma_0, \xi_0) = (0.5, 0)$			
Bias	0.005	0.002	0.000	-0.005	-0.001	-0.008
SEE	0.169	0.181	0.199	0.121	0.129	0.142
SSE	0.174	0.185	0.205	0.124	0.134	0.145
CP	0.943	0.950	0.946	0.942	0.943	0.949
			$(\gamma_0, \xi_0) = (0.5, 0.2)$			
Bias	0.017	0.033	0.012	0.020	0.024	0.018
SEE	0.169	0.177	0.196	0.120	0.127	0.139
SSE	0.171	0.183	0.199	0.123	0.128	0.142
CP	0.940	0.937	0.952	0.938	0.946	0.945

Table 2. Estimation results for θ with the link function $g(x) = x$

	$n = 100$			$n = 200$		
θ_0	0	0.2	0.5	0	0.2	0.5
			$(\gamma_0, \xi_0) = (0, 0)$			
Bias	0.008	0.005	-0.006	0.009	0.004	0.009
SEE	0.269	0.261	0.249	0.191	0.186	0.178
SSE	0.287	0.277	0.246	0.201	0.187	0.185
CP	0.932	0.928	0.948	0.939	0.953	0.940
			$(\gamma_0, \xi_0) = (0, 0.2)$			
Bias	0.035	0.041	0.047	0.042	0.036	0.040
SEE	0.259	0.254	0.245	0.186	0.181	0.174
SSE	0.282	0.265	0.257	0.191	0.184	0.184
CP	0.927	0.934	0.924	0.929	0.936	0.921
			$(\gamma_0, \xi_0) = (0.5, 0)$			
Bias	-0.007	0.015	0.010	0.001	0.011	0.010
SEE	0.247	0.239	0.233	0.176	0.172	0.166
SSE	0.256	0.259	0.249	0.180	0.179	0.177
CP	0.939	0.930	0.927	0.939	0.933	0.936
			$(\gamma_0, \xi_0) = (0.5, 0.2)$			
Bias	0.052	0.051	0.045	0.040	0.051	0.042
SEE	0.244	0.239	0.231	0.174	0.171	0.166
SSE	0.252	0.258	0.237	0.178	0.171	0.169
CP	0.932	0.917	0.935	0.929	0.947	0.930

To generate $Y_i(T_{i,j})$ at each observation time $T_{i,j}$, we considered

$$E\{Y_i(T_{i,j})|Z_i, \mathbf{b}_i(t)\} = g\{\mu_0(t)e^{\theta Z_i}\}b_{1i}(t),$$

and obtained $Y_i(T_{i,j})$ by first generating $Y_i^*(T_{i,j})$ from a Poisson distribution with the mean function of $Y_i^*(t)$ being equal to $g\{\mu_0(t)e^{\theta Z_i}\}b_{1i}(t)E\{I(t \le C_i)|Z_i, \mathcal{B}_{it}\}$, and then taking $Y_i(T_{i,j}) = \frac{Y_i^*(T_{i,j})}{E\{I(T_{i,j} \le C_i)|Z_i, \mathcal{B}_{it}\}}$. The results given below are based on the sample sizes of 100 and 200 with 1,000 replications and $W(t) = 1$.

We took $\lambda_0 = 2$, $d\Lambda_0(t) = \frac{5}{t}(e^{0.5} - e^{-0.5})(e^t - e^{-t})dt$, $b_{1i} = \frac{2e^{v_i}}{e-1/e}$, $b_{2i}(t) = \frac{2te^{u_i+v_it}}{(e^{0.5}-e^{-0.5})(e^t-e^{-t})}$ and $b_{3i} = v_i$ with u_i and v_i being random numbers generated from uniform distributions over $(-0.5, 0.5)$ and $(-1, 1)$, respectively. Table 1 shows the estimation results for θ based on the simulated data with the link function $g(x) = \log(x)$, $\mu_0(t)\} = e^{2t}$, and the true values of (γ, ξ) being equal to (0, 0), (0, 0.2), (0.5, 0), (0.5, 0.2). The table includes the estimated bias given by the average of the proposed estimators $\hat{\theta}$ minus the true value θ_0, the average of the estimated standard errors (SEE), the empirical sampling standard error (SSE) and the 95% empirical coverage probability (CP). It can be seen that the proposed approach seems to perform well. Specifically, the proposed estimate seems to be unbiased and the estimated standard errors agree well with the empirical ones. Also as expected, the CP's are close to their nominal levels and the standard errors become smaller when sample sizes increase.

In addition to the scenarios presented by Table 1, we investigated those with various link functions and random effects. For example, the results given in Table 2 were obtained with the same setups as those for Table 1 except that $g(x) = x$ and $\mu_0(t) = 2t$. Such results all suggest that the proposed procedure perform well for practical situations. To further study how various link functions affected the estimation results, we also calculated the averaged sum of absolute residuals ($\overline{RES}$) for each scenario, defined as

$$\overline{RES} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}|d\widehat{M}_i(T_{i,j})|.$$

Table 3. Averaged sum of residuals based on results from Tables 1 and 2 when $n = 200$

	$g(t) = \log(t)$			$g(t) = t$		
θ_0	0	0.2	0.5	0	0.2	0.5
$(\gamma_0, \xi_0) = (0,0)$	3.134	3.584	4.260	3.116	3.473	4.147
$(\gamma_0, \xi_0) = (0, 0.2)$	3.311	3.795	4.525	3.288	3.678	4.435
$(\gamma_0, \xi_0) = (0.5, 0)$	4.115	4.872	5.982	4.117	4.658	5.736
$(\gamma_0, \xi_0) = (0.5, 0.2)$	4.404	5.177	6.367	4.379	5.056	6.255

Table 3 presents the results obtained for scenarios represented by Tables 1 and 2 when $n = 200$, where the baseline mean function is common for $Y_i(t)$ given $\mathbf{b}_i(t)$ and Z_i. The results show that when the choice of $g(\cdot)$ is reasonable, such residuals are comparable whether the covariate effects are additive (for $g(x) = \log(x)$) or multiplicative (for $g(x) = x$) to the response process.

One question of practical interest is that for longitudinal data when the observation process is informative, whether

Table 4. Estimation results of θ based on the proposed procedure and the one given by Zhao et al. (2013), when $g(x) = log(x)$ and $\xi_0 = 0$

θ_0	$n = 100$			$n = 200$		
	$BIAS$	$BIAS_1^*$	$BIAS_2^*$	BIAS	$BIAS_1^*$	$BIAS_2^*$
$\gamma_0 = 0.5$						
0	0.008	-0.140	-0.118	0.000	-0.148	-0.135
0.2	0.003	-0.162	-0.141	-0.002	-0.154	-0.143
0.5	-0.009	-0.167	-0.139	-0.011	-0.169	-0.149
$\gamma_0 = 0.8$						
0	-0.007	-0.208	-0.196	0.000	-0.212	-0.181
0.2	-0.005	-0.210	-0.202	-0.002	-0.224	-0.198
0.5	-0.009	-0.220	-0.192	-0.011	-0.246	-0.195

Table 5. Analysis results for the skin cancer data

	Est.	SEE	90% CI	p-value
γ_1	0.529	0.072	(0.410, 0.648)	< 0.001
γ_2	0.566	0.072	(0.448, 0.684)	< 0.001
ξ_1	1.203	0.171	(0.922, 1.484)	< 0.001
ξ_2	1.038	0.171	(0.757, 1.319)	< 0.001
		$g(x) = x$		
θ_1	-0.448	0.187	(-0.814, -0.082)	0.017
θ_2	1.164	0.225	(0.723, 1.064)	< 0.001
		$g(x) = \log(x)$		
θ_1	-0.225	0.123	(-0.427, -0.024)	0.066
θ_2	0.972	0.118	(0.777, 1.167)	< 0.001

some existing procedure applies to the situations as considered by models (1)–(3). While there are limited procedures for regression analysis based on a class of transformation models for the response process, most of them model possible correlation between $Y_i(t)$ and $\widetilde{N}_i(t)$ by incorporating a specific function of $\widetilde{N}_i(s)$, $s \le t$ to the marginal mean of $Y_i(t)$ (Sun et al., 2005; Li et al., 2013; Zhao et al., 2013), for example, a function denoted by $h(\cdot)$ in Zhao et al. (2013). One possible drawback is that such applications are highly subject to the specific form of $h(\cdot)$, which cannot capture correlations of an arbitrary form. To illustrate this numerically, we considered both the proposed estimation procedure and the one given in Zhao et al. (2013). Note that the latter also considered a possible dependent terminal event time D_i but assumed a noninformative C_i given Z_i, For ease of comparison, we made $D_i > C_i$ in our scenarios and used each subject's last observation time as C_i when applied the competing procedure. Table 4 presents the estimation results for θ obtained for $g(x) = \log(x)$, $b_{1i} = \frac{1}{2}(\exp\{0.5 - v_i\} - \exp\{-0.5 - v_i\} + G_i)$, $b_{2i}(t) = \frac{(t+1)\exp\{v_i(t+1)\}}{e^{0.5(t+1)} - e^{-0.5(t+1)}}$, $b_{3i} = v_i$, $d\Lambda_0(t) = \frac{20t}{t+1}\{e^{0.5(t+1)} - e^{-0.5(t+1)}\}$, $\mu_0(t) = \exp\{5t\}$, with v_i and G_i being random numbers from the uniform distribution over $(-0.5, 0.5)$ and the gamma distribution with mean 1 and variance 0.5, respectively. In the table, $BIAS$ represents the estimated bias from the proposed estimate; $BIAS_1^*$ and $BIAS_2^*$ denote the estimated biases given by Zhao et al. (2013) using $h(\mathcal{F}_{it}) = \widetilde{N}(t-)$ and $h(\mathcal{F}_{it}) = 0$, respectively. The results suggest that the proposed estimates still appear to be unbiased, but the competing method could give substantially biased estimates for θ when the correlations between $Y_i(t)$, $N_i(t)$ and C_i introduced by $\mathbf{b}_i(t)$ are misinterpreted by $h(\cdot)$ or totally ignored.

6. AN APPLICATION

In this section, we applied the proposed methodology described in the previous sections to longitudinal data arising from a skin cancer study conducted by the University of Wisconsin Comprehensive Cancer Center in Madison, Wisconsin (Li et al., 2011; Zhang et al., 2013). One main objective of this double-blind, placebo-controlled randomized Phase III clinical trial is to evaluate the effectiveness of $0.5g/m^2$/day PO difluoromethylornithine (DFMO) in reducing the recurrence rates of basal cell carcinoma (BCC) for patients with a history of skin cancers. At each visit, the numbers of BCC occurrences since the previous visit were recorded. Each patient was scheduled to be assessed every six months; however as expected, the actual observation times vary from patient to patient. Besides a patient's treatment group (placebo or DFMO), the study also provided information on the number of prior skin cancer occurrences which is shown to be significantly related to the skin cancer recurrence process. For the analysis, we focus on the 290 patients with at least one observation. Among them, 161 patients had one or two skin cancer occurrences prior to the study, and the others had experienced more.

In the following, we consider covariates defined by $\mathbf{Z}_i = (Z_{i1},\ Z_{i2})'$, where $Z_{i1} = 1$ if patient i was given the DFMO treatment and $Z_{i1} = 0$ otherwise, and $Z_{i2} = 1$ if the patient had experienced more than two (up to 35) skin cancer occurrences and $Z_{i2} = 0$ if not, $i = 1, \ldots, 290$. $Y_i(t)$ represents the total number of BCC occurrences observed up to time t. The longest follow-up time was scaled to be $\tau = 1$, which corresponds to 1,879 days in the original data set.

To apply the proposed estimation procedure, we assumed that the skin cancer recurrence process, the observation process and the hazard of censoring can be described by models (1)–(3), respectively. Following the notation above, the primary interest is to estimate θ_1, the effect of DFMO. Table 5 presents the analysis results obtained by applying the proposed estimation procedure with $W(t) = 1$. We considered two link functions: $g(x) = x$ and $g(x) = \log(x)$, and the results include the point estimates (Est.), the estimated standard errors (SEE), the estimated 90% confidence intervals (CI's) and p-values for tests with the null hypotheses assuming no covariate effects. At the significance level of $\alpha = 0.1$, the results suggest that DFMO has significantly reduced the recurrence rates of BCC, and a more severe skin cancer history appears to be positively correlated with the recurrence rate of skin cancer. Such results appear consistent with those concluded by Li et al. (2014) for both choices of link

functions. In addition, the results also suggest that both the observation and follow-up times significantly depend on the covariates.

To assess the adequacy of our models above, we applied the goodness-of-fit test derived in Section 4 and obtained the p-values of 0.801 and 0.383, respectively, for $g(x) = x$ and $g(x) = \log(x)$. This suggests that while both of our link functions appear to be reasonable for the data, the former is preferred over the latter.

7. CONCLUDING REMARKS

This paper considers regression analysis of longitudinal data when both the observation and follow-up times may be informative about the underlying response process of interest. For the problem, we present a class of semiparametric transformation models for the response process which allow possible correlations to be characterized by time-dependent random effects. Comparing with existing models that assume either independence or structured dependence based on fixed forms or distributions, the proposed models provide flexibility for modeling both the underlying response process and its correlation to other processes. For parameter estimation, an easy-to-implement estimating equation approach is developed and both finite and asymptotic properties of the resulting estimators are established. In addition, the extensive simulation study indicated that the approach works well for practical situations and the approach is applied to a skin cancer study which motivated the research.

We note several possible directions for future work. First for simplicity, we assumed that the dependence between $Y_i(t)$ and $N_i(t)$ in models (1)–(2) can be completely characterized by random effects $\mathbf{b}_i(t)$ and covariates $\mathbf{Z}_i(t)$. However in practice, one may want to incorporate more terms to the content of $g(\cdot)$ as well when additional information is available. For example, if it is known from pivotal trials or experiences that a longitudinal response depends on the length of period since subject i is last observed, one may consider modifying model (1) as follows:

$$E\{Y_i(t)|\mathbf{Z}_i(t), \mathbf{b}_i(t)\} = g\{\mu_0(t)e^{\theta'\mathbf{Z}_i(t)+\alpha(t-T_{i,j})}\}b_{1i}(t),$$

where $j = max\{k : T_{i,k} \le t\}$ and $T_{i,j}$ represents subject i's last observation time. In such cases, the same methodology immediately applies for estimating θ and α together, by replacing θ and $\mathbf{Z}_i(t)$ by $(\theta', \alpha')'$ and $\big(\mathbf{Z}_i(t)', t - T_{i,j}\big)'$, respectively, in the estimation procedure. Second, the focus of the article has been on regression analysis of the response process $Y_i(t)$, therefore, $\mathbf{b}_i(t)$ was treated as a shared latent vector. However, if one is solely interested in calculating any correlation between $Y_i(t)$, $N_i(t)$ and C_i at certain times, one may usually need a distribution assumption on $\mathbf{b}_i(t)$ and apply some existing procedures for inference (Lipsitz et al., 2002; He et al., 2009; Sun et al., 2007, 2007b; Li et al., 2013). Other than the effects of $\mathbf{b}_i(t)$, we have assumed the proportional rates and additive hazards models, respectively, on $N_i(t)$ and C_i. In context of dependent processes, a procedure that is robust to such models is another interesting direction for future research.

ACKNOWLEDGEMENTS

The authors wish to thank the editor, the associate editor and the two reviewers for their constructive comments and suggestions that led to a great improvement of this manuscript. This work was partially supported by funds provided by National Science Foundation (grant DMS-1208978 to Sun), National Institutes of Health (grant 2 R37 AI054165 to Sun) and The University of North Carolina at Charlotte (to Sun and FRG 1-11172 to Li).

APPENDIX A

Proof of Theorem 1

To derive the asymptotic properties of the proposed estimator $\hat{\theta}$, we need the following regularity conditions.

(C1). $\{\widetilde{N}_i(\cdot), Y_i(\cdot), C_i, \mathbf{Z}_i(\cdot)\}_{i=1}^n$ are independent and identically distributed.

(C2). There exists a $\tau > 0$ such that $P(C_i \ge \tau) > 0$.

(C3). Both $\widetilde{N}_i(t)$ and $Y_i(t)$ $(0 \le t \le \tau,\ i = 1, \ldots, n)$ are bounded.

(C4). $W(t)$ and $\mathbf{Z}_i(\cdot)$, $i = 1, \ldots, n$, have bounded variations and $W(t)$ converges almost surely to a deterministic function $w(t)$ uniformly in $t \in [0, \tau]$.

(C5). $A_\theta = E\int_0^\tau W(t)\dot{g}\{\mu_0(t)e^{\theta_0'\mathbf{Z}_i(t)}\}\{\mathbf{Z}_i(t) - e_z(t)\}^{\otimes 2}e^{\theta_0'\mathbf{Z}_i(t)+\eta_0'\mathbf{X}_i^*(t)}\mu_0(t)d\Lambda_1^*(t)$ and $\Omega_\eta = E\Big[\int_0^\tau \{\mathbf{X}_i^*(t) - \bar{x}^*(t)\}^{\otimes 2}e^{\eta_0'\mathbf{X}_i^*(t)}d\Lambda_1^*(t)\Big]$ are both positive definite.

Define

$$U_1(\theta; \hat{\eta}) = \sum_{i=1}^n \int_0^\tau W(t)\mathbf{Z}_i(t) \times \left[Y_i(t)d\widetilde{N}_i(t) - e^{\hat{\eta}'\mathbf{X}_i^*(t)}g\{\hat{\mu}_0(t)e^{\theta'\mathbf{Z}_i(t)}\}d\hat{\Lambda}_1^*(t)\right] = 0$$

and note that $\hat{\mu}_0(t)$ satisfies

$$\sum_{i=1}^n \left[Y_i(t)d\widetilde{N}_i(t) - e^{\hat{\eta}'\mathbf{X}_i^*(t)}g\{\hat{\mu}_0(t)e^{\theta'\mathbf{Z}_i(t)}\}d\hat{\Lambda}_1^*(t)\right] = 0,\ 0 \le t \le \tau. \quad (A.1)$$

Let

$$\hat{A}_\theta(\theta) = -n^{-1}\partial U_1(\theta, \hat{\eta})/\partial\theta', \hat{A}_\eta(\eta) = -n^{-1}\partial U_1(\theta_0, \eta)/\partial\eta',$$
$$A_\theta = \lim_{n\to\infty} \hat{A}_\theta(\theta_0) \text{ and } A_\eta = \lim_{n\to\infty} \hat{A}_\eta(\eta_0).$$

The consistency of $\hat{\theta}$ and $\hat{\eta}$ follows from the facts that $U_1(\theta_0; \hat{\eta})$ and $U_\eta(\eta_0)$ both tend to 0 in probability as $n \to \infty$, and that $\hat{A}_\theta(\theta)$ and $-n^{-1}\partial U_\eta(\eta)/\partial\eta$ both converge uniformly to the positive definite matrices A_θ and Ω_η over θ and η, respectively, in neighborhoods around the true values

θ_0 and η_0. Then the Taylor series expansions of $U_1(\hat{\theta};\hat{\eta})$ at $(\theta_0;\hat{\eta})$ and (θ_0,η_0) yield $n^{1/2}(\hat{\theta}-\theta_0) = A_\theta^{-1}n^{-1/2}U_1(\theta_0;\hat{\eta}) + o_p(1) = A_\theta^{-1}\Big\{n^{-1/2}U_1(\theta_0;\eta_0) - A_\eta n^{1/2}(\hat{\eta}-\eta_0)\Big\} + o_p(1)$. The proof of Theorem 1 is sketched as follows:

(1) First, using some derivation operation to $U_1(\theta;\hat{\eta})$ and (A.1), we can get

$$\begin{aligned}\hat{A}_\theta(\theta) &= n^{-1}\sum_{i=1}^n \int_0^\tau W(t) g\{\hat{\mu}_0(t)e^{\hat{\theta}'\mathbf{Z}_i(t)}\} \\ &\times \big\{\mathbf{Z}_i(t) - \hat{E}_Z(t)\big\}^{\otimes 2} e^{\theta'\mathbf{Z}_i(t)+\hat{\eta}'\mathbf{X}_i^*(t)} d\hat{\Lambda}_1^*(t).\end{aligned}$$

(2) The use of Taylor expansions of $U_1(\theta_0;\eta_0)$ and (A.1) at $\mu_0(t)$ yield

$$\begin{aligned}U_1(\theta_0;\eta_0) &= \sum_{i=1}^n \int_0^\tau w(t)\Big(\mathbf{Z}_i(t) - e_z(t)\Big) dM_i(t) \\ &- \sum_{i=1}^n \int_0^\tau w(t)\Big(\mathbf{Z}_i(t) - e_z(t)\Big) g\{\mu_0(t)e^{\theta_0'\mathbf{Z}_i(t)}\} \\ &\times e^{\eta_0'\mathbf{X}_i^*(t)} d\{\hat{\Lambda}_1^*(t;\eta_0) - \Lambda_1^*(t)\} + o_p(n^{1/2}).\end{aligned}$$

It follows from (7) that

$$\hat{\Lambda}_1^*(t;\eta_0) - \Lambda_1^*(t) = \frac{1}{n}\sum_{i=1}^n \int_0^t \frac{dM_i^*(t)}{s^{(0)}(t)} + o_p(n^{-1/2}).$$

Thus

$$U_1(\theta_0;\eta_0) = \sum_{i=1}^n (\xi_{1i} - \xi_{2i}) + o_p(n^{1/2}), \tag{15}$$

where $\xi_{1i} = \int_0^\tau w(t)\Big(\mathbf{Z}_i(t) - e_z(t)\Big)dM_i(t)$, $\xi_{2i} = \int_0^\tau \frac{w(t)d(t)}{s^{(0)}(t)} dM_i^*(t)$ and $d(t) = \lim_{n\to\infty}\hat{D}(t;\theta_0,\eta_0)$.

(3) Differentiation of $U_1(\theta_0,\eta)$ and (A.1) with respect to η' yields

$$\begin{aligned}\hat{A}_\eta(\eta) = n^{-1}\sum_{i=1}^n \int_0^\tau & W(t) g\{\hat{\mu}_0(t)e^{\theta_0'\mathbf{Z}_i(t)}\} e^{\eta'\mathbf{X}_i^*(t)} \\ & \times \{\mathbf{Z}_i(t) - \hat{E}_Z(t)\}\{\mathbf{X}_i^*(t) - \bar{X}^*(t;\eta)\}' d\hat{\Lambda}_1^*(t;\eta).\end{aligned}$$

(4) According to equation (6) and the arguments similar to Lin et al. (2000), one can show that

$$n^{1/2}\{\hat{\eta}-\eta_0\} = \Omega_\eta^{-1} n^{-1/2}\sum_{i=1}^n \zeta_i + o_p(1) \tag{A.2}$$

where $\Omega_\eta = E\Big[\int_0^\tau \big\{\mathbf{X}_i^*(t) - \bar{x}^*(t)\big\}^{\otimes 2} e^{\eta_0'\mathbf{X}_i^*(t)} d\Lambda_1^*(t)\Big]$ and $\zeta_i = \int_0^\tau \Big(\mathbf{X}_i^*(t) - \bar{x}^*(t)\Big) dM_i^*(t)$.

Combining the results in steps (1)–(4), we have

$$U_1(\theta_0;\hat{\eta}) = \sum_{i=1}^n (\xi_{1i} - \xi_{2i} - \xi_{3i}) + o_p(n^{1/2}),$$

and hence

$$\sqrt{n}(\hat{\theta}-\theta_0) = A_\theta^{-1} n^{-1/2}\sum_{i=1}^n (\xi_{1i} - \xi_{2i} - \xi_{3i}) + o_p(1), \tag{A.3}$$

where $\xi_{3i} = \int_0^\tau A_\eta \Omega_\eta^{-1}\big\{\mathbf{X}_i^*(t) - \bar{x}^*(t)\big\} dM_i^*(t)$. Then it follows from the multivariate central limit theorem that the conclusions hold.

APPENDIX B

Proof of the null distribution of $\mathcal{F}(t,z)$

Define $V(\hat{\theta},\hat{\eta}) = \sum_{i=1}^n \int_0^t I(\mathbf{Z}_i(s) \le z) d\widehat{M}_i(s)$. By applying the Taylor expansion,

$$\begin{aligned}\mathcal{F}(t,x;\hat{\theta},\hat{\eta}) &= n^{-1/2}V(\theta_0,\eta_0) + \frac{\partial V(\theta_0,\eta_0)}{n\partial\eta'}\sqrt{n}(\hat{\eta}-\eta_0) \\ &+ \frac{\partial V(\theta_0,\hat{\eta})}{n\partial\theta'}\sqrt{n}(\hat{\theta}-\theta_0) + o_p(1).\end{aligned}$$

By following arguments and manipulations similar to those in Appendix A, it can be shown

$$V(\theta_0,\eta_0) = \sum_{i=1}^n \{u_{1i}(t,z) - u_{2i}(t,z)\} + o_p(n^{1/2}),$$

where $u_{1i}(t,z) = \int_0^t \{I\big(\mathbf{Z}_i(s) \le z\big) - e_I(s,z)\} dM_i(s)$, $u_{2i}(t,z) = \int_0^t \frac{\Gamma(s)}{s^{(0)}(s)} dM_i^*(s)$ and $\Gamma(t) = \lim_{n\to\infty}\hat{\Gamma}(t;\theta_0,\eta_0)$.

Also $\frac{\partial V(\theta_0,\eta_0)}{n\partial\eta'}$ and $\frac{\partial V(\theta_0,\hat{\eta})}{n\partial\theta'}$ can be estimated by $-\hat{V}_\eta(t,z)$ and $-\hat{V}_\theta(t,z)$, respectively. In addition, we obtained

$$n^{1/2}\{\hat{\eta}-\eta_0\} = \Omega_\eta^{-1} n^{-1/2}\sum_{i=1}^n \zeta_i + o_p(1)$$

and

$$\sqrt{n}(\hat{\theta}-\theta_0) = A_\theta^{-1} n^{-1/2}\sum_{i=1}^n (\xi_{1i} - \xi_{2i} - \xi_{3i}) + o_p(1),$$

from (A.2) and (A.3). Therefore, $\mathcal{F}(t,z;\hat{\theta},\hat{\eta})$ can be expressed as a sum of i.i.d. mean-zero terms for fixed t. By the multivariate central limit theorem, $\mathcal{F}(t,z)$ converges in finite-dimensional distribution to a mean-zero Gaussian distribution. Since $\mathcal{F}(t,z)$ is tight based on the empirical process theory, $\mathcal{F}(t,z)$ converges weakly to a mean-zero Gaussian process that can be approximated by $\widehat{\mathcal{F}}(t,z)$ given by equation (11).

Received 7 October 2014

REFERENCES

DIGGLE, P. J., LIANG, K. Y. and ZEGER, S. L. (1994). *The Analysis of Longitudinal Data*. Oxford University Press, Oxford.

CHENG, S. C. and WEI, L. J. (2000). Inferences for a semiparametric model with panel data. *Biometrika* **87** 89–97. MR1766830

He, X., Tong, X. and Sun, J. (2009). Semiparametric analysis of panel count data with correlated observation and follow-up times. *Lifetime Data Analysis* **15** 177–196. MR2510863
Hu, X. J., Sun, J. and Wei, L. J. (2003). Regression parameter estimation from panel counts. *Scandinavian Journal of Statistics* **30** 25–43. MR1963891
Huang, C. Y., Wang, M. C. and Zhang, Y. (2006). Analysing panel count data with informative observation times. *Biometrika* **93** 763–775. MR2285070
Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, New York. MR1924807
Li, N., Sun, L. and Sun, J. (2010). Semiparametric transformation models for panel count data with dependent observation process. *Statistics in Biosciences* **2** (2) 191–210.
Li, N., Zhao, H. and Sun, J. (2013). Semiparametric transformation models for panel count data with correlated observation and follow-up times. *Statistics in Medicine* **32** (17) 3039–3054. MR3073834
Li, Y., Zhao, H., Sun, J. and Kim, K. M. (2014). Nonparametric tests for panel count data with unequal observation processes. *Computational Statistics & Data Analysis* **73** 103–111. MR3147977
Lin, D. Y., Oaks, D. and Ying, Z. (1998). Additive hazards regression with current status data. *Biometirka* **85** (2) 289–298. MR1649115
Lin, D. Y., Wei, L. J., Yang, I. and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society, Series B* **62** 711–730. MR1796287
Lin, D. Y. and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of American Statistical Association* **96** 103–126. MR1952726
Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Gelber, R., and Lipshultz, S. (2002). Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics* **58** 621–630. MR1933535
Sun, J. and Kalbfleisch, J. D. (1995). Estimation of the mean function of point processes based on panel count data. *Statistica Sinica* **5** 279–289. MR1329298
Sun, J., Park, D-H., Sun, L. and Zhao, X. (2005). Semiparametric regression analysis of longitudinal data with informative observation times. *Journal of American Statistical Association* **100** 882–889. MR2201016
Sun, J., Sun, L. and Liu, D. (2007). Regression analysis of longitudinal data in the presence of informative observation and censoring times. *Journal of the American Statistical Association* **102** 1397–1406. MR2372540
Sun, J., Tong, X. and He, X. (2007b). Regression analysis of panel count data with dependent observation times. *Biometrics* **63** 1053–1059. MR2414582
Sun, J. and Wei, L. J. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *Journal of the Royal Statistical Society, Series B* **62** 293–302. MR1749540
Sun, L., Song, X., Zhou, J. and Liu, L. (2013). Joint analysis of longitudinal data with informative observation times and a dependent terminal event. *Journal of the American Statistical Association* **107** (498) 688–700. MR2980077
Sun, Y. (2010). Estimation of semiparametric regression model with longitudinal data. *Lifetime Data Analysis* **16** (2) 271–298. MR2608289
Wellner, J. A. and Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Annals of Statistics* **28** 779–814. MR1792787
Wellner, J. A. and Zhang, Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *Annals of Statistics* **35** 2106–2142. MR2363965
Welsh, A. H., Lin, X. and Carroll, R. J. (2002). Marginal longitudinal nonparametric regression: locality and efficiency of spline and kernel methods. *Journal of American Statistical Association* **97** 482–493. MR1941465
Zhang, Y. (2002). A semiparametric pseudolikelihood estimation method for panel count data. *Biometrika* **89** 39–48. MR1888344
Zhang, Z., Sun, J. and Sun, L. (2005). Statistical analysis of current status data with informative observation times. *Statistics in Medicine* **24** 1399–1407. MR2134566
Zhao, H., Li, Y. and Sun, J. (2013). Analyzing panel count data with dependent observation process and a terminal event. *The Canadian Journal of Statistics* **41** (1) 174–191. MR3030791
Zhao, X., Balakrishnan, N. and Sun, J. (2011). Nonparametric inference based on panel count data. *Test* **20** 1–42. MR2806303
Zhao, X., Zhou, J. and Sun, L. (2011), Semiparametric transformation models with time-varying coefficients for recurrent and terminal events. *Biometrics* **67** 404–414. MR2829009
Zhao, X. and Tong, X. (2011). Semiparametric regression analysis of panel count data with informative observation times. *Computational Statistics and Data Analysis* **55** (1) 291–300. MR2736555
Zhao, X., Tong, X. and Sun, J. (2013b). Robust estimation for panel count data with informative observation times. *Computational Statistics and Data Analysis* **57** 33–40. MR2981070

Yang Li
Department of Mathematics and Statistics
UNC Charlotte
Charlotte, NC 28223
USA
E-mail address: Y.Li@uncc.edu

Yanqing Sun
Department of Mathematics and Statistics
UNC Charlotte
Charlotte, NC 28223
USA
E-mail address: yasun@uncc.edu