Published in final edited form as:

Biometrics. 2012 December ; 68(4): 1093–1102. doi:10.1111/j.1541-0420.2012.01794.x.

Time-Varying Latent Effect Model for Longitudinal Data with Informative Observation Times

Na Cai*, Wenbin Lu**, and Hao Helen Zhang***

Department of Statistics, North Carolina State University, Raleigh, North Carolina, U.S.A

Summary

In analysis of longitudinal data, it is not uncommon that observation times of repeated measurements are subject-specific and correlated with underlying longitudinal outcomes. Taking account of the dependence between observation times and longitudinal outcomes is critical under these situations to assure the validity of statistical inference. In this article, we propose a flexible joint model for longitudinal data analysis in the presence of informative observation times. In particular, the new procedure considers the shared random-effect model and assume a time-varying coefficient for the latent variable, allowing a flexible way of modeling longitudinal outcomes while adjusting their association with observation times. Estimating equations are developed for parameter estimation. We show that the resulting estimators are consistent and asymptotically normal, with variance-covariance matrix that has a closed form and can be consistently estimated by the usual plug-in method. One additional advantage of the procedure is that, it provides a unified framework to test whether the effect of the latent variable is zero, constant, or time-varying. Simulation studies show that the proposed approach is appropriate for practical use. An application to a bladder cancer data is also given to illustrate the methodology.

Keywords

Estimating equation method; Informative observation times; Longitudinal data analysis; Time-varying effect

1. Introduction

Longitudinal data arise in many medical follow-up studies, where the primary response of interest is repeatedly measured over time. Various methods have been proposed for analysis of longitudinal data, including the marginal approach based on estimating equations (e.g. Liang and Zeger, 1986; Robins et al., 1995; Yi and Cook, 2002; Welsh et al., 2002) and the subject-specific approach via random effects (e.g. Laird and Ware, 1982; Wu and Carroll, 1988; Follmann and Wu, 1995; Wulfsohn and Tsiatis, 1997; Zeng and Cai, 2005). Most of these methods assume that observation times of longitudinal measurements are predetermined and the same across study subjects or independent of underlying longitudinal outcomes.

However, in many longitudinal studies, subjects' observation times can be very irregular and possibly subject-specific, and the frequency and timing of subjects' visits may be correlated

*ncai@ncsu.edu.

**lu@stat.ncsu.edu

***hzhang2@ncsu.edu

7. Supplementary Materials

The Web Appendix referenced in Sections 3 and 5 is available with this paper at the Biometrics website on Wiley Online Library.

with underlying longitudinal outcomes. One motivating example is the bladder cancer study (Sun and Wei, 2000; Sun et al., 2005), where patients may have multiple tumor recurrences and repeated clinical visits to remove newly occurred tumors. Here the longitudinal outcomes of interest are the number of newly occurred tumors between visits. In this study, the number of clinical visits has a large variation among the patients, as some patients have significantly more visits than others, suggesting a possible correlation between observation times and tumor recurrence. A number of methods have been developed for handling irregular subject-specific observation times and studying the correlation between longitudinal outcomes and associated observation times. For example, Lin and Ying (2001) proposed a semiparametric marginal regression model for longitudinal data, where a flexible counting process is used to model subject-specific observation times. In their model, it assumed that subjects' observation times are independent of underlying longitudinal outcomes given covariates. To allow more flexible dependence between observation times and longitudinal outcomes, Lin et al. (2004) proposed a class of inverse intensity of visit process-weighted estimators by allowing the intensity function of the visiting process to depend on the past observed longitudinal outcomes. Similar idea was also studied by Buzkova and Lumley (2007) for generalized linear models. Sun et al. (2005) extended the semiparametric marginal regression model of Lin and Ying (2001) by including a function of the past observed visiting history as a predictor. In addition, various approaches have been proposed to use subject-specific random effects to describe the association between longitudinal outcomes and observations times. The joint likelihood based estimation (Ryu et al., 2007; Liu et al., 2008; Liu and Huang, 2009) and estimating equations based methods (Huang et al., 2006; Sun et al., 2007; Liang et al., 2009) were developed accordingly. All of these methods assume that the effect of latent variables on longitudinal outcomes is time-invariant. However, since both observation times and longitudinal outcomes are time-dependent stochastic processes, it is more plausible to assume that the effect of latent variable on longitudinal outcomes may also change over time.

In this article, we propose a semiparametric model with time-varying coefficient for the latent variable in analysis of longitudinal data with informative observation times. This new method provides a flexible way of modeling the latent variable effect on the longitudinal outcome process while adjusting its association with the observation time process. Moreover, it naturally generalizes the semiparametric marginal regression model of Lin and Ying (2001) and provides a broad class of goodness-of-fit tests for testing the time-varying effect against null and constant effects of the latent variable.

The rest of the paper is organized as follows. Section 2 introduces the joint model and specifies its assumptions. Section 3 presents a set of unbiased estimating equations for parameter estimation and the asymptotic properties of the resulting estimators. In Section 4, we propose a class of goodness-of-fit test statistics for testing different hypotheses on the effect of the latent variable including the null and time-invariant effects, and develop a resampling method to compute the associated p -values of the tests. Section 5 is devoted to simulation studies and the analysis of a bladder cancer data. Some discussions are given in Section 6. The theoretical proofs are given in the Appendix.

2. Notation and Model Specification

Consider a longitudinal study of n subjects. For the i th subject, let $Y_i(t)$ be the underlying longitudinal outcome at time t and X_i be the $p \times 1$ vector of baseline covariates. In addition, let $T_{i1} < T_{i2} < \dots$ denote the potential observation times of longitudinal outcomes and C_i denote the follow-up time. Define $N_i(t) = N_i^*(t \wedge C_i)$, where $N_i^*(t) = \sum_{j=1}^{\infty} I(T_{ij} \leq t)$ and $t \wedge C_i$

$\equiv \min(t, C_i)$. Then $Y_i(t)$ is only observed at the jump points of $N_i(t)$. Let $m_i \equiv N_i(\infty)$ be the total number of observations made on the i th subject

We consider the following time-varying latent effect model for the longitudinal outcome:

$$E\{Y_i(t)|X_i, V_i\} = \alpha_{10}(t) + V_i \alpha_{20}(t) + \beta_0' X_i, \quad t > 0, \quad (1)$$

where $\alpha_{10}(t)$ and $\alpha_{20}(t)$ are unspecified smooth functions, the random effect V_i is a latent variable with $E(V_i|X_i) = 1$ and $\text{Var}(V_i) > 0$, and β_0 is a $p \times 1$ vector of unknown regression coefficients. The assumption $E(V_i|X_i) = 1$ is needed only for model identifiability, as $\alpha_{20}(t)$ is a nonparametric function. Model (1) characterizes the conditional mean of $Y_i(\cdot)$, and does not specify its distribution and the dependence structure among repeated measurements. Given the covariates X_i and the latent variable V_i , the intensity of the observation process $N_i^*(t)$ is modeled via a multiplicative frailty model (Andersen et al., 1993), i.e.

$$\lambda(t|X_i, V_i) = V_i \lambda_0(t) \exp(\gamma_0' X_i), \quad (2)$$

where $\lambda_0(t)$ is an unspecified baseline intensity function and γ_0 is a $p \times 1$ vector of unknown regression coefficients.

Two key features of models (1) and (2) are: first, the dependence between $Y_i(t)$ and $N_i^*(t)$ is characterized by a shared latent variable V_i ; second, V_i has a time-varying effect on the longitudinal outcome process $Y_i(t)$. Throughout this article, we make the following assumptions: (i) $Y_i(\cdot)$ is independent of $N_i^*(\cdot)$ given X_i and V_i ; (ii) C_i is independent of $Y_i(\cdot)$, $N_i^*(\cdot)$ and V_i given X_i . In models (1) and (2), the distribution of V_i , the functions $\lambda_0(t)$, $\alpha_{10}(t)$ and $\alpha_{20}(t)$ are all left unspecified, so they are treated as nonparametric components.

3. Estimation Procedure and Asymptotics

Note that model (2) is a special case of the proportional mean/rate model, i.e.

$E\{dN_i^*(t)|X_i\} = \exp(\gamma_0' X_i) \lambda_0(t) dt$. Therefore, the parameters γ_0 and $\Lambda_0(t) \equiv \int_0^t \lambda_0(u) du$ can be consistently estimated using the equations proposed by Lin et al. (2000). Let $\hat{\gamma}$ and $\hat{\Lambda}_0(\cdot)$ denote the resulting estimators of γ_0 and $\Lambda_0(\cdot)$, respectively.

Next, we propose estimating equations for the parameters in model (1). Since $N_i^*(\cdot)$ is a Poisson process given X_i and V_i , we have $E\{dN_i(t)|m_i, V_i, X_i, C_i\} = I(C_i \geq t) \frac{m_i}{\Lambda_0(C_i)} d\Lambda_0(t)$. Based on model (1) and our assumptions, it follows that

$$E[\{Y_i(t) - \beta_0' X_i\} dN_i(t) | m_i, V_i, X_i, C_i] = \{\alpha_{10}(t) + V_i \alpha_{20}(t)\} \xi_i(t) \frac{m_i}{\Lambda_0(C_i)} d\Lambda_0(t),$$

where $\xi_i(t) = I(C_i \geq t)$. This motivates us to consider the cumulative effects

$\mathcal{A}_{10}(t) = \int_0^t \alpha_{10}(s) d\Lambda_0(s)$ and $\mathcal{A}_{20}(t) = \int_0^t \alpha_{20}(s) d\Lambda_0(s)$ as similarly studied in Lin and Ying (2001) and define

$$M_i^*(t; \beta, \Lambda, \mathcal{A}_1, \mathcal{A}_2) = \int_0^t \left[\{Y_i(s) - \beta' X_i\} dN_i(s) - \xi_i(s) \frac{m_i}{\Lambda(C_i)} d\mathcal{A}_1(s) - \xi_i(s) \frac{m_i V_i}{\Lambda(C_i)} d\mathcal{A}_2(s) \right].$$

It is easy to show that $E\{M_i^*(t; \beta_0, \Lambda_0, \mathcal{A}_{10}, \mathcal{A}_{20}) | m_i, V_i, X_i, C_i\} = 0$ for any t . If the latent variables V_i 's were known, we can estimate $\mathcal{A}_{10}(t)$, $\mathcal{A}_{20}(t)$ and β_0 simultaneously using the following estimating equations:

$$\sum_{i=1}^n \begin{pmatrix} 1 \\ m_i \end{pmatrix} dM_i^*(t; \beta, \Lambda, \mathcal{A}_1, \mathcal{A}_2) = 0, \quad \forall t \in [0, \tau], \quad (3)$$

$$\sum_{i=1}^n \int_0^\tau X_i dM_i^*(t; \beta, \Lambda, \mathcal{A}_1, \mathcal{A}_2) = 0, \quad (4)$$

where τ is the maximum follow-up time of the study. The solutions $\mathcal{A}_1^*(t; \beta, \Lambda)$ and $\mathcal{A}_2^*(t; \beta, \Lambda)$ of equations (3) are given by the following matrix operation

$$\begin{pmatrix} \mathcal{A}_1^*(t; \beta, \Lambda) \\ \mathcal{A}_2^*(t; \beta, \Lambda) \end{pmatrix} = \int_0^t \begin{pmatrix} \sum_{i=1}^n \xi_i(s) \frac{m_i}{\Lambda(C_i)} & \sum_{i=1}^n \xi_i(s) \frac{m_i V_i}{\Lambda(C_i)} \\ \sum_{i=1}^n \xi_i(s) \frac{m_i^2}{\Lambda(C_i)} & \sum_{i=1}^n \xi_i(s) \frac{m_i^2 V_i}{\Lambda(C_i)} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n \{Y_i(s) - \beta' X_i\} dN_i(s) \\ \sum_{i=1}^n m_i \{Y_i(s) - \beta' X_i\} dN_i(s) \end{pmatrix}.$$

Plugging $\mathcal{A}_1^*(\cdot; \beta, \Lambda)$ and $\mathcal{A}_2^*(\cdot; \beta, \Lambda)$ into equation (4), we have

$$U^*(\beta; \Lambda) = \sum_{i=1}^n \int_0^\tau \left[X_i - B^*(t; \Lambda) \{A^*(t; \Lambda)\}^{-1} \begin{pmatrix} 1 \\ m_i \end{pmatrix} \right] \{Y_i(t) - \beta' X_i\} dN_i(t) = 0, \quad (5)$$

where the matrices $A^*(t; \Lambda)$ and $B^*(t; \Lambda)$ are defined as

$$A^*(t; \Lambda) = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n \xi_i(t) \frac{m_i}{\Lambda(C_i)} & \frac{1}{n} \sum_{i=1}^n \xi_i(t) \frac{m_i V_i}{\Lambda(C_i)} \\ \frac{1}{n} \sum_{i=1}^n \xi_i(t) \frac{m_i^2}{\Lambda(C_i)} & \frac{1}{n} \sum_{i=1}^n \xi_i(t) \frac{m_i^2 V_i}{\Lambda(C_i)} \end{pmatrix},$$

and

$$B^*(t; \Lambda) = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n \xi_i(t) \frac{m_i}{\Lambda(C_i)} X_i & \frac{1}{n} \sum_{i=1}^n \xi_i(t) \frac{m_i V_i}{\Lambda(C_i)} X_i \end{pmatrix}.$$

In practice, the latent variables V_i 's are not observed. Therefore, equation (5) is not computable based on the observed data and we need to replace $A^*(t; \Lambda)$ and $B^*(t; \Lambda)$ in (5) by their estimable counterparts that have the same limits. Note that given X_i , V_i and C_i , the total number of observations m_i follows a Poisson distribution with the mean equal to

$V_i \Lambda_0(C_i) e^{\gamma_0' X_i}$. It follows that

$$E(m_i | V_i, X_i, C_i) = V_i \Lambda_0(C_i) e^{\gamma_0' X_i}, \quad E\{(m_i^2 - m_i) | V_i, X_i, C_i\} = V_i^2 \{\Lambda_0(C_i) e^{\gamma_0' X_i}\}^2 \quad \text{and}$$

$$E\{(m_i^3 - 3m_i^2 + 2m_i) | V_i, X_i, C_i\} = V_i^3 \{\Lambda_0(C_i) e^{\gamma_0' X_i}\}^3. \quad \text{Define}$$

$$A(t; \gamma, \Lambda) = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n \xi_i(t) \frac{m_i}{\Lambda(C_i)} & \frac{1}{n} \sum_{i=1}^n \xi_i(t) \frac{m_i^2 - m_i}{\Lambda^2(C_i) e^{\gamma' X_i}} \\ \frac{1}{n} \sum_{i=1}^n \xi_i(t) \frac{m_i^2}{\Lambda(C_i)} & \frac{1}{n} \sum_{i=1}^n \xi_i(t) \frac{m_i(m_i - 1)^2}{\Lambda^2(C_i) e^{\gamma' X_i}} \end{pmatrix},$$

and

$$B(t; \gamma, \Lambda) = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n \xi_i(t) \frac{m_i}{\Lambda(C_i)} X_i & \frac{1}{n} \sum_{i=1}^n \xi_i(t) \frac{m_i^2 - m_i}{\Lambda^2(C_i) e^{\gamma' X_i}} X_i \end{pmatrix}.$$

By simple algebra, we can show that $A(t; \gamma_0, \Lambda_0)$ and $B(t; \gamma_0, \Lambda_0)$ have the same limits as $A^*(t, \Lambda_0)$ and $B^*(t, \Lambda_0)$. Replacing $A^*(t, \Lambda)$ and $B^*(t, \Lambda)$ by $A(t; \gamma, \Lambda)$ and $B(t; \gamma, \Lambda)$ respectively, the estimators of $\mathcal{A}_1(t)$ and $\mathcal{A}_2(t)$, given β , γ and Λ , are given by

$$\begin{pmatrix} \mathcal{A}_1(t; \beta, \gamma, \Lambda) \\ \mathcal{A}_2(t; \beta, \gamma, \Lambda) \end{pmatrix} = \int_0^t \{A(s; \gamma, \Lambda)\}^{-1} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n \{Y_i(s) - \beta' X_i\} dN_i(s) \\ \frac{1}{n} \sum_{i=1}^n m_i \{Y_i(s) - \beta' X_i\} dN_i(s) \end{pmatrix}, \quad (6)$$

and the estimating function for β can be written as

$$\tilde{U}(\beta; \gamma, \Lambda) = \sum_{i=1}^n \int_0^\tau \left\{ X_i - B(t; \gamma, \Lambda) \{A(t; \gamma, \Lambda)\}^{-1} \begin{pmatrix} 1 \\ m_i \end{pmatrix} \right\} \{Y_i(t) - \beta' X_i\} dN_i(t). \quad (7)$$

In addition, it can be shown that the following function

$$\frac{1}{n} \sum_{i=1}^n \int_0^\tau \left\{ X_i - B(t; \gamma_0, \Lambda_0) \{A(t; \gamma_0, \Lambda_0)\}^{-1} \begin{pmatrix} 1 \\ m_i \end{pmatrix} \right\} g(t) dN_i(t)$$

converges to zero in probability for any function $g(\cdot)$. Then the estimating function $\tilde{U}(\beta; \gamma, \Lambda)$ can be extended to the following class of estimating functions

$$\tilde{U}_g(\beta; \gamma, \Lambda) = \sum_{i=1}^n \int_0^\tau \left\{ X_i - B(t; \gamma, \Lambda) \{A(t; \gamma, \Lambda)\}^{-1} \begin{pmatrix} 1 \\ m_i \end{pmatrix} \right\} \{Y_i(t) - \beta' X_i - g(t)\} dN_i(t).$$

Following Lin and Ying (2001), to reduce the variance of $\tilde{U}_g(\beta; \gamma, \Lambda)$, a good choice of $g(t)$ could be $\bar{Y}(t) - \beta' \bar{Y}(t)$, where $\bar{Y}(t) = \sum_{i=1}^n \xi_i(t) \frac{m_i}{\Lambda(C_i)} Y_i(t) / \sum_{i=1}^n \xi_i(t) \frac{m_i}{\Lambda(C_i)}$ and $\bar{X}(t) = \sum_{i=1}^n \xi_i(t) \frac{m_i}{\Lambda(C_i)} X_i / \sum_{i=1}^n \xi_i(t) \frac{m_i}{\Lambda(C_i)}$. In general, the value of $\bar{Y}(t)$ may not be evaluable at all the observation times, but it can be approximated by $\bar{Y}^*(t) = \sum_{i=1}^n \xi_i(t) \frac{m_i}{\Lambda(C_i)} Y_i^*(t) / \sum_{i=1}^n \xi_i(t) \frac{m_i}{\Lambda(C_i)}$, where $Y_i^*(t)$ is the observed longitudinal outcome at the time point nearest to t . Setting $g(t) = \bar{Y}^*(t) - \beta' \bar{Y}(t)$ and plugging $\hat{\gamma}$ and $\hat{\Lambda}_0$ into $\tilde{U}_g(\beta; \gamma, \Lambda)$, we obtain the estimating function $U(\beta; \hat{\gamma}, \hat{\Lambda}_0)$ of β , defined by

$$\sum_{i=1}^n \int_0^\tau \left\{ X_i - B(t; \hat{\gamma}, \hat{\Lambda}_0) \{A(t; \hat{\gamma}, \hat{\Lambda}_0)\}^{-1} \begin{pmatrix} 1 \\ m_i \end{pmatrix} \right\} [\{Y_i(t) - \bar{Y}^*(t)\} - \beta' \{X_i - \bar{X}(t)\}] dN_i(t).$$

Let $\hat{\beta}$ denote the solution to equation $U(\beta; \hat{\gamma}, \hat{\Lambda}_0) = 0$. Then we have

$$\begin{aligned} \hat{\beta} = & \left[\sum_{i=1}^n \int_0^\tau \left\{ X_i - B(t; \hat{\gamma}, \hat{\Lambda}_0) \{A(t; \hat{\gamma}, \hat{\Lambda}_0)\}^{-1} \begin{pmatrix} 1 \\ m_i \end{pmatrix} \right\} \{X_i - \bar{X}(t)\}' dN_i(t) \right]^{-1} \\ & \times \sum_{i=1}^n \int_0^\tau \left\{ X_i - B(t; \hat{\gamma}, \hat{\Lambda}_0) \{A(t; \hat{\gamma}, \hat{\Lambda}_0)\}^{-1} \begin{pmatrix} 1 \\ m_i \end{pmatrix} \right\} \{Y_i(t) - \bar{Y}^*(t)\} dN_i(t). \end{aligned}$$

Moreover, the estimators of $\mathcal{A}_{10}(t)$ and $\mathcal{A}_{20}(t)$ are given by $\mathcal{A}_1(t; \hat{\beta}, \hat{\gamma}, \hat{\Lambda}_0)$ and $\mathcal{A}_2(t; \hat{\beta}, \hat{\gamma}, \hat{\Lambda}_0)$, respectively. The asymptotic properties of $\hat{\beta}$, $\mathcal{A}_1(\cdot; \hat{\beta}, \hat{\gamma}, \hat{\Lambda}_0)$ and $\mathcal{A}_2(\cdot; \hat{\beta}, \hat{\gamma}, \hat{\Lambda}_0)$ are established in the following theorems.

Theorem 1—Under conditions (C1)–(C4) given in the Appendix, we have $\|\hat{\beta} - \beta_0\| \rightarrow 0$, $\sup_{t \in [0, \tau]} |\mathcal{A}_1(t; \hat{\beta}, \hat{\gamma}, \hat{\Lambda}_0) - \mathcal{A}_{10}(t)| \rightarrow 0$ and $\sup_{t \in [0, \tau]} |\mathcal{A}_2(t; \hat{\beta}, \hat{\gamma}, \hat{\Lambda}_0) - \mathcal{A}_{20}(t)| \rightarrow 0$ in probability as $n \rightarrow \infty$.

Theorem 2—Under conditions (C1)–(C4) given in the Appendix, as n goes to infinity, $n^{1/2}(\hat{\beta} - \beta_0)$ converges in distribution to a mean-zero normal random vector with variance-covariance matrix of $D^{-1}\Sigma(D')^{-1}$.

Theorem 3—Under conditions (C1)–(C4) given in the Appendix, as n goes to infinity,

$$n^{1/2} \begin{pmatrix} \mathcal{A}_1(t; \hat{\beta}, \hat{\gamma}, \hat{\Lambda}_0) - \mathcal{A}_{10}(t) \\ \mathcal{A}_2(t; \hat{\beta}, \hat{\gamma}, \hat{\Lambda}_0) - \mathcal{A}_{20}(t) \end{pmatrix} \text{ converge weakly to a mean-zero Gaussian process.}$$

The technical derivation of the above asymptotic results becomes challenging due to the increased complexity of the estimating equations. The proofs of the above theorems and the definitions of the matrices D and Σ are given in the Web Appendix. Consistent estimators of D and Σ can be obtained by the usual plug-in method. For example, a consistent estimator of D is given by

$$\hat{D} = \frac{1}{n} \sum_{i=1}^n \int_0^\tau \left\{ X_i - B(t; \hat{\gamma}, \hat{\Lambda}_0) \{A(t; \hat{\gamma}, \hat{\Lambda}_0)\}^{-1} \begin{pmatrix} 1 \\ m_i \end{pmatrix} \right\} \{X_i - \bar{X}(t)\}' dN_i(t).$$

4. Model Diagnostics

In practice, it is of interest to test whether the effect of the latent variable on longitudinal outcomes is zero (no effect) or constant (time-invariant effect), i.e., to test the null hypotheses: $H_{01}: a_{20}(t) \equiv 0$ and $H_{02}: a_{20}(t) \equiv \theta$, where θ is an unspecified constant. The first test provides a diagnostic for the marginal model of Lin and Ying (2001), checking the independence assumption between longitudinal outcomes and observation times given covariates, while the second one tests the general structure of the latent variable effect. In this section, we develop a class of goodness-of-fit tests for these null hypotheses.

As given in the Appendix, we have the following asymptotic representations:

$$n^{1/2}\{\widehat{\Lambda}_0(t) - \Lambda_0(t)\} = n^{-1/2} \sum_{i=1}^n \varphi_i(t) + o_p(1) \text{ and}$$

$$n^{1/2} \begin{pmatrix} \mathcal{A}_1(t; \widehat{\beta}, \widehat{\gamma}, \widehat{\Lambda}_0) - \mathcal{A}_{10}(t) \\ \mathcal{A}_2(t; \widehat{\beta}, \widehat{\gamma}, \widehat{\Lambda}_0) - \mathcal{A}_{20}(t) \end{pmatrix} = n^{-1/2} \sum_{i=1}^n \Phi_i(t; \beta_0, \gamma_0, \Lambda_0) + o_p(1)$$

for $t \in [0, \tau_1]$. Under H_{10} , we have $\mathcal{A}_{20}(t) \equiv 0$, while $\mathcal{A}_{20}(t) \equiv \theta \Lambda_0(t)$ under H_{20} . The test statistic for H_{10} can be constructed as

$$T_{n,1} = \sup_{t \in [\tau_1, \tau_2]} |n^{1/2} \mathcal{A}_2(t; \widehat{\beta}, \widehat{\gamma}, \widehat{\Lambda}_0)|,$$

where $0 < \tau_1 < \tau_2 < \tau$. Here the interval $[\tau_1, \tau_2]$ denotes the region over which time period one would like to test whether the latent variable effect is time-dependent or constant. In general, we can take $\tau_1 = 0$, while τ_2 is chosen to avoid the instability of the estimates at the right tail due to few recurrent observation times. To obtain the critical values of the test statistic, we use the resampling method of Lin et al. (1994). Specifically, let $\widehat{\varphi}_i(t)$ and $\widehat{\Phi}_{i2}(t)$ denote the consistent estimators of $\varphi_i(t)$ and the second component of $\Phi_i(t; \beta_0, \gamma_0, \Lambda_0)$, obtained by the usual plug-in method. We construct the perturbed test statistics

$T_{n,1}^* = \sup_{t \in [\tau_1, \tau_2]} |n^{-1/2} \sum_{i=1}^n \widehat{\Phi}_{i2}(t) Z_i|$, where $Z_1, \dots, Z_n$ are iid standard normal random variables. It can be shown that the asymptotic distribution of $T_{n,1}^*$ given data is the same as that of $T_{n,1}$. Therefore, by generating a large set of variables Z_i^* s, we can use the sample quantiles of $T_{n,1}^*$ to approximate the critical values of $T_{n,1}$. The associated p -value can then be computed.

To construct a test statistic for H_{20} , we need to estimate θ under the null. One possible estimator can be obtained as $\widehat{\theta} = \int_{\tau_1}^{\tau_2} w(t) \{\mathcal{A}_2(t; \widehat{\beta}, \widehat{\gamma}, \widehat{\Lambda}_0) / \widehat{\Lambda}_0(t)\} dt / \int_{\tau_1}^{\tau_2} w(t) dt$, where $w(t)$ is a positive weight function. For example, we can set $w(t) = 1$ when there is a jump of $\sum_{i=1}^n N_i(t)$ at time point t , and 0 otherwise. Then $\int_{\tau_1}^{\tau_2} w(t) dt$ counts the number of unique observation times on the interval $[\tau_1, \tau_2]$ among all study subjects. A test statistic is then given by $T_{n,\theta} = \sup_{t \in [\tau_1, \tau_2]} |n^{1/2} \{\mathcal{A}_2(t; \widehat{\beta}, \widehat{\gamma}, \widehat{\Lambda}_0) - \widehat{\theta} \widehat{\Lambda}_0(t)\}|$. To obtain the perturbed test statistics using the resampling method, define

$$\widehat{\theta}^* = \int_{\tau_1}^{\tau_2} w(s) n^{-1/2} \sum_{i=1}^n \{\widehat{\Phi}_{i2}(s) - \widehat{\theta} \widehat{\varphi}_i(s)\} Z_i \frac{ds}{\widehat{\Lambda}_0(s)} / \int_{\tau_1}^{\tau_2} w(s) ds.$$

Then the perturbed test statistic is given by

$$T_{n,\theta}^* = \sup_{t \in [\tau_1, \tau_2]} |n^{-1/2} \sum_{i=1}^n \{\widehat{\Phi}_{i2}(t) - \widehat{\theta} \widehat{\varphi}_i(t)\} Z_i - \widehat{\theta}^* \widehat{\Lambda}_0(t)|.$$

Note that the quantities $\hat{\Phi}_{\mathcal{Z}}(t)$, $\hat{\varphi}_{\mathcal{A}}(t)$ and $\hat{\Lambda}_0(t)$ in $T_{n,\theta}^*$ are all step functions with the jumps only at the observed recurrence times. Therefore, the supreme only needs to be taken over all the observed recurrence times in the interval $[\tau_1, \tau_2]$. The p -value of the test can be computed accordingly based on the perturbed test statistics.

5. Numerical Studies

5.1 Simulation Studies

Simulations are conducted to evaluate the performance of the proposed estimation procedure given in Section 3. Longitudinal outcomes and observation times are generated from model (1) and (2), respectively. In models (1) and (2), two covariates $X_i = (X_{i1}, X_{i2})'$ are considered, where X_{i1} is generated from a Bernoulli distribution with the success probability of 0.5 and X_{i2} is from a normal distribution $N(0, 0.5^2)$. The longitudinal outcome $Y_{\mathcal{A}}(t)$ is given by

$$Y_i(t) = \alpha_{10}(t) + V_i \alpha_{20}(t) + \beta_0' X_i + \varepsilon_i(t),$$

where the measurement error $\varepsilon_i(t)$ is generated from $N(0, 0.5)$ and independent of X_i and V_i . We set $\beta_0' = (\beta_{01}, \beta_{02}) = (1, -1)$, $\alpha_{10}(t) = 1 + \sin(t)$ and consider four scenarios for $\alpha_{20}(t)$: $\alpha_{20}(t) = 0$, 1, $1/(1+t)$ or t . The latent variable V_i is generated from a mixture of two random variables and its distribution depends on the binary covariate X_{i1} . Specifically, we set $V_i = (1 - X_{i1})U_{i1} + X_{i1}U_{i2}$, where U_{i1} and U_{i2} are independent positive random variables with mean of 1. We consider two situations:

- a. *mixed gamma frailty*: U_{i1} is generated from a gamma distribution with mean of 1 and variance of 0.5 and U_{i2} is from a gamma distribution with mean of 1 and variance of 0.8;
- b. *mixed log normal frailty*: U_{i1} is generated from a log-normal distribution with mean of 1 and variance of 0.5 and U_{i2} is from a log-normal distribution with mean of 1 and variance of 0.8.

In both cases, the marginal distribution of V_i has mean of 1 and variance of 0.65. Moreover, in model (2), we choose $\gamma_0 = (\gamma_{01}, \gamma_{02}) = (0.5, -0.5)$ and $\lambda_0(t) \equiv 2$. The censoring time C_i is generated from a uniform distribution on the interval $[1, 4]$, which gives the average number of observations per subject around 6.8.

Under each scenario, we consider 500 replications with sample size of $n = 200$. The simulation results for our proposed estimators of β_0 are summarized in Table 1. For comparison, we also report the results for the estimators of Lin and Ying (2001), which assumes independence between longitudinal outcomes and observation times given covariates. The results show that our estimators are nearly unbiased for all the scenarios, the mean of estimated standard error (SE) obtained using the plug-in method is close to the sample standard deviation (SD) of parameter estimates, and the empirical coverage probability of 95% Wald-type confidence interval (CI) is close to the nominal level. For the estimators of Lin and Ying (2001), we make the following observations. When $\alpha_{20}(t) = 0$, i.e., the independence assumption of Lin and Ying (2001) is correct, the estimators are essentially unbiased and the coverage probability of the CIs is close to 95%. However, in all the other scenarios with $\alpha_{20}(t) \neq 0$, the parameter estimates of β_{01} exhibit some biases, and the amount of bias can be very large in some situations. For example, for the case of mixture gamma frailty and $\alpha_{20}(t) = t$, the bias is 0.39. In addition, due to the misspecification of correlation between the outcome and observation time processes, the procedure tends to

cause an inflated sample standard deviation for the parameter estimates. For example, for the case of mixture log-normal frailty and $\alpha_{20}(t) = t$, the SDs of the estimates for β_{01} and β_{02} are respectively 0.877 and 1.024, while their corresponding SEs are 0.536 and 0.518. In summary, the proposed estimators give a comparable performance as Lin and Ying (2001) when $\alpha_2(t) = 0$ (the case where outcomes are independent of observation times given covariates). When the latent effect is not zero or time-dependent, the new estimators perform much better in terms of their negligible biases and much smaller SDs.

Next we investigate the performance of our proposed estimators for the nonparametric functions $\mathcal{A}_{10}(\cdot)$ and $\mathcal{A}_{20}(\cdot)$. In Figures 1 and 2, we plot the mean of estimated nonparametric functions $\mathcal{A}_1(\cdot; \hat{\beta}, \hat{\gamma}, \hat{\Lambda}_0)$ and $\mathcal{A}_2(\cdot; \hat{\beta}, \hat{\gamma}, \hat{\Lambda}_0)$ and their associated pointwise 95% confidence intervals for the mixture gamma and mixture log-normal distributions of the latent variable, respectively. In both Figures, the left panel is for the function $\mathcal{A}_{10}(\cdot)$ while the right panel is for $\mathcal{A}_{20}(\cdot)$. From the plots, we can see that our proposed estimators of nonparametric functions are nearly unbiased and their associated pointwise 95% confidence intervals cover the underlying true functions for all the scenarios.

5.2 Analysis of Bladder Cancer Data

We apply our proposed method to the bladder cancer data studied by Sun and Wei (2000) and Zhang (2002). The data consists of 85 patients with superficial bladder tumors, where 47 patients were randomly assigned to the placebo group and 38 to the thiotepa treatment group. For each patient, besides the treatment indicator and initial number of tumors, the times (in months) of clinical visits and number of newly observed tumors since the last visit of the patient were also recorded. The maximum of follow-up time is 53 months. The frequency of clinical visits varies a lot, ranging from 1 to 38. The average number of clinical visits is 13.5 (11.5) for the treatment group and 8.7 (4.7) for the control group. The main interest is to study the effects of treatment and number of initial tumors on tumor recurrence rate. This dataset has been studied by many authors, including Sun et al. (2005, 2007), Liang et al. (2009) and Buzkova and Lumley (2010). It was found that the visiting times are correlated with the number of newly occurring tumors. In particular, Sun et al. (2007) and Liang et al. (2009) used latent variables to describe their association. However, the effect of the latent variable on longitudinal outcomes are all assumed time-invariant. Here we analyze the data using the proposed time-varying latent effect model.

As in Sun et al. (2007) and Liang et al. (2009), for patient i , let $Y_i(t)$ be the logarithm of the number of newly occurring tumors at time t since the last visit plus 1 to avoid 0. We consider two covariates: X_{i1} denotes the treatment indicator (1 for treatment and 0 for control) and X_{i2} denotes the logarithm of the initial number of tumors plus 1. The results of our estimators are summarized in Table 2. For comparison, we also report the results for the estimators of Lin and Ying (2001). The results indicate that the treatment indicator has a significant effect on the visiting process while the number of initial tumors does not. In addition, the patients in the treatment group tend to have more clinical visits than those in the control group ($\hat{\gamma}_1 = 0.616$ with the estimated standard error of 0.155), which agrees with the empirical frequencies of clinical visits in the two groups. For the regression parameters in the time-varying latent effect model, our results imply that the treatment indicator and initial number of tumors both have significant effects on the number of tumor recurrences, where the treatment has the beneficial effect, i.e. reducing the number of newly occurring tumors, while the number of initial tumor has the detrimental effect. Similar results were also obtained for Lin and Ying's estimators. However, a main difference is that the estimated treatment effect from Lin and Ying's method is substantially elevated comparing to ours (our estimator: $\hat{\beta}_1 = -0.152$ with the estimated standard error of 0.042; Lin and Ying's estimator: $\hat{\beta}_1 = -0.192$ with the estimated standard error of 0.057). A similar finding was also observed by Liang et al (2009). One possible interpretation is that since the patients

in the treatment group have more clinical visits, their tumors were treated more often and as a result, they tend to have less newly occurring tumors compared to those in the control group. Therefore, ignoring the latent association between the visiting process and longitudinal outcomes, the estimated treatment effect may be biased as we have found in simulations.

To investigate the effect of the latent variable on longitudinal outcomes, we plot the estimated functions $\hat{\Lambda}_0(t)$, $\hat{\mathcal{A}}_1(t; \hat{\beta}, \hat{\gamma}, \hat{\Lambda}_0)$ and $\hat{\mathcal{A}}_2(t; \hat{\beta}, \hat{\gamma}, \hat{\Lambda}_0)$ and their corresponding pointwise 95% confidence intervals in Figures 3 and 4, respectively. Based on these plots, we make the following observations. First, the zero line is not contained in the pointwise 95% confidence intervals of $\hat{\mathcal{A}}_{20}(t)$ (see the right panel of Figure 4), which suggests that the function $a_{20}(t)$ is significantly different from zero and thus the visiting process is correlated with longitudinal outcomes via the latent variable; Second, the estimated function $\hat{\mathcal{A}}_2(t; \hat{\beta}, \hat{\gamma}, \hat{\Lambda}_0)$ is clearly not proportional to the estimated function $\hat{\Lambda}_0(t)$, which implies that the effect of the latent variable on longitudinal outcomes (i.e. $a_{20}(t)$) is not a constant since

$\mathcal{A}_{20}(t) = \int_0^t a_{20}(s) d\Lambda_0(s)$. Third, comparing the estimated curves of $\hat{\mathcal{A}}_2(t; \hat{\beta}, \hat{\gamma}, \hat{\Lambda}_0)$ and $\hat{\Lambda}_0(t)$, it can be seen that the function $a_{20}(t)$ is negative for small t , but its magnitude diminishes and eventually it becomes positive as t increases. To further justify our findings, we perform the goodness-of-fit tests proposed in Section 4 for the null hypotheses: $H_{01}: a_{20}(t) \equiv 0$ (no effect) and $H_{02}: a_{20}(t) \equiv \theta$ (time-invariant effect). For each test, we set $\tau_1 = 0$ and $\tau_2 = 26$, and conduct 1000 resamplings for computing the critical values. The resulting p -values are 0.05 and 0.06, respectively, which imply that the latent variable has a marginally significant effect on longitudinal outcomes but its effect may be time-varying. To further elaborate this, we repeat the second test on a shorter interval $[\tau_1, \tau_2] = [15, 26]$. The estimated constant effect $\hat{\theta} = -0.415$ with the p -value of 0.046. This confirms that the latent variable effect is truly time-varying.

Finally, we also compared our method with the parametric joint likelihood approach proposed by Liu et al. (2008), in which the random effects and error term in the longitudinal model are assumed to follow independent normal distributions. The results are given in the Web Appendix. In summary, the results based on Liu et al. (2008)'s model generally agree with the findings from our method, especially for the estimation of β_0 . In addition, the estimate for the association parameter θ is -0.104 with the standard error of 0.095, which also supports the negative association between the observation times and longitudinal outcomes. One possible reason for such a negative association is that the more often the patients visited the clinic and received the treatment, the less chance they will have tumor recurrence. The association based on Liu et al. (2008)'s model is not statistically significant, which might be partly due to the time-varying effect of the latent variable on the longitudinal outcomes.

6. Discussion

The proposed time-varying latent effect model provides a flexible way for analysis of longitudinal data with informative observation times and a unified framework for estimating and testing the complex effect of the latent variable on longitudinal outcomes. The proposed estimating equations are relatively simple to implement and allow a rigorous development of asymptotic normality with an explicit formula for the variance-covariance matrix. The computation of the proposed estimators has been implemented using R codes, which are available from authors upon request.

In this paper, we consider only time-independent covariates with constant coefficients, but our method can be easily extended to accommodate time-dependent covariates with time-varying coefficients following the discussions of Lin and Ying (2001). In addition, we

assume that the censoring or drop-out time is noninformative, i.e. independent of longitudinal outcomes and observations times given covariates, which may be restrictive for practical applications and can be relaxed as in Sun et al. (2007) using the method of Wang et al. (2001) for recurrent event data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the associate editor and two referees for their comments that substantially improved the presentation of the article. We also thank Professor Lei Liu for providing us the SAS codes to implement the joint likelihood estimation method used in the real data application. WL's work was partially supported by NIH/NCI grants R01 CA140632 and R01 CA051962. HZ's work was partially supported by NIH/NCI grants P01 CA142538 and R01 CA085848, and NSF grant DMS-0645293.

References

- Andersen, PK.; Borgan, O.; Gill, RD.; Keiding, N. Statistical Models Based on Counting Processes. New York: Springer-Verlag; 1993.
- Buzkova P, Lumley T. Longitudinal data analysis for generalized linear models with follow-up dependent on outcome-related variables. *The Canadian Journal of Statistics*. 2007; 35:485–500.
- Follmann D, Wu M. An approximate generalized linear model with random effects for informative missing data. *Biometrics*. 1995; 51:151–168. [PubMed: 7766771]
- Huang CY, Wang MC, Zhang Y. Analysing panel count data with informative observation times. *Biometrika*. 2006; 93:763–775.
- Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics*. 1982; 38:963–974. [PubMed: 7168798]
- Liang KY, Zeger SL. Longitudinal data-analysis using generalized linear-models. *Biometrika*. 1986; 73:13–22.
- Liang Y, Lu W, Ying Z. Joint modeling and analysis of longitudinal data with informative observation times. *Biometrics*. 2009; 65:377–384. [PubMed: 18759841]
- Lin DY, Fleming TR, Wei LJ. Confidence bands for survival curves under the proportional hazards model. *Biometrika*. 1994; 81:73–81.
- Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society, Series B*. 2000; 62:711–730.
- Lin DY, Ying Z. Semiparametric and nonparametric regression analysis of longitudinal data (with discussion). *Journal of the American Statistical Association*. 2001; 96:103–112.
- Lin H, Scharfstein DO, Rosenheck RA. Analysis of longitudinal data with irregular outcome-dependent follow-up. *Journal of the Royal Statistical Society, Series B*. 2004; 66:791–813.
- Liu L, Huang X. Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. *Journal of the Royal Statistical Society, Series C*. 2009; 58:65–81.
- Liu L, Huang X, O'Quigley J. Analysis of longitudinal data in the presence of informative observation times and a dependent terminal event, with application to medical cost data. *Biometrics*. 2008; 64:950–958. [PubMed: 18162110]
- Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression-models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*. 1995:90.
- Ryu D, Sinha D, Mallick B, Lipsitz SR, Lipshultz SE. Longitudinal studies with outcome-dependent follow-up: Models and Bayesian regression. *Journal of the American Statistical Association*. 2007; 102:952–961. [PubMed: 18392118]

- Sun J, Park DH, Sun L, Zhao X. Semiparametric regression analysis of longitudinal data with informative observation times. *Journal of the American Statistical Association*. 2005; 100:882–889.
- Sun J, Sun L, Liu D. Regression analysis of longitudinal data in the presence of informative observation and censoring times. *Journal of the American Statistical Association*. 2007; 102:1379–1406.
- Sun J, Wei LJ. Regression analysis of panel count data with covariate-dependent observation and censoring times. *Journal of the Royal Statistical Society, Series B*. 2000; 62:293–302.
- Wang MC, Qin J, Chiang CT. Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association*. 2001; 96:1057–1065.
- Welsh AH, Lin X, Carroll RJ. Marginal longitudinal nonparametric regression: Locality and efficiency of spline and kernel methods. *Journal of the American Statistical Association*. 2002; 97:482–493.
- Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*. 1988; 44:175–188.
- Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. *Biometrics*. 1997; 53:330–339. [PubMed: 9147598]
- Yi GY, Cook RJ. Marginal methods for incomplete longitudinal data arising in clusters. *Journal of the American Statistical Association*. 2002; 97:1071–1080.
- Zeng D, Cai J. Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *Annals of Statistics*. 2005; 33:2132–2163.
- Zhang Y. A semiparametric pseudolikelihood estimation method for panel count data. *Biometrika*. 2002; 89:39–48.

Figure 1. Mixture gamma distribution: solid lines are underlying true functions; dashed lines are averages of estimated functions; dash-dot lines are pointwise 95% confidence intervals

Figure 2. Mixture log-normal distribution: solid lines are underlying true functions; dashed lines are averages of estimated functions; dash-dot lines are pointwise 95% confidence intervals

Figure 3.
The solid line represents the point estimates of $\Lambda_0(t)$ and dashed lines are corresponding pointwise 95% confidence intervals.

Figure 4.

The solid lines represent the point estimates of $A_{10}(t)$ (left panel) and $A_{20}(t)$ (right panel). Dashed lines are corresponding pointwise 95% confidence intervals.

Table 1

Summary of simulation results for mixture gamma frailty and mixture log normal frailty

$\alpha_{20}(\ell)$	β_0	mixture gamma frailty							
		LY method				Our method			
		Est.	SD	SE	CP	Est.	SD	SE	CP
0	β_{01}	1.001	0.044	0.040	0.924	1.001	0.045	0.050	0.950
	β_{02}	-1.001	0.041	0.039	0.946	-1.001	0.041	0.043	0.948
$1/(1+\theta)$	β_{01}	1.139	0.138	0.127	0.800	1.003	0.118	0.115	0.940
	β_{02}	-1.001	0.169	0.130	0.896	-1.009	0.140	0.130	0.960
1	β_{01}	1.282	0.281	0.251	0.794	1.006	0.222	0.213	0.936
	β_{02}	-1.005	0.344	0.256	0.888	-1.018	0.267	0.244	0.954
ℓ	β_{01}	1.390	0.439	0.383	0.820	1.011	0.306	0.315	0.934
	β_{02}	-1.011	0.517	0.384	0.894	-1.022	0.354	0.333	0.944

$\alpha_{20}(\ell)$	β_0	mixture log normal frailty							
		LY method				Our method			
		Est.	SD	SE	CP	Est.	SD	SE	CP
0	β_{01}	0.988	0.040	0.040	0.948	1.000	0.055	0.056	0.960
	β_{02}	-1.000	0.042	0.039	0.930	-1.002	0.055	0.049	0.956
$1/(1+\theta)$	β_{01}	1.135	0.234	0.173	0.874	1.006	0.125	0.120	0.940
	β_{02}	-0.985	0.299	0.172	0.892	-1.008	0.160	0.130	0.946
1	β_{01}	1.283	0.507	0.352	0.868	1.017	0.235	0.221	0.944
	β_{02}	-0.971	0.630	0.346	0.910	-1.015	0.304	0.245	0.940
ℓ	β_{01}	1.411	0.877	0.536	0.880	1.035	0.329	0.340	0.942
	β_{02}	-0.967	1.024	0.518	0.916	-1.018	0.419	0.349	0.948

[†]Est., mean of parameter estimates; SD, sample standard deviation of parameter estimates; SE, mean of estimated standard errors; CP, empirical coverage probabilities for 95% confidence intervals.

\$watermark-text

\$watermark-text

\$watermark-text

Table 2

Analysis of the bladder cancer data

	β					
	γ		LY Method		Our Method	
	Est.	SE	Est.	SE	Est.	SE
Treatment	0.616	0.155	-0.192	0.057	-0.152	0.042
Number of initial tumor	-0.062	0.178	0.190	0.054	0.205	0.050

[#]Est., parameter estimates; SE, estimated standard errors.