Analysis of Panel Count Data with Time-Dependent Covariates and Informative Observation Process

Sha Fang[1], Haixiang Zhang[2*,3], Liuquan Sun[3], Dehui Wang[2]

[1]School of Mathematical Sciences, Peking University, Beijing 100871, China

[2*]Mathematics School of Jilin University, Changchun 130012, China

[3]Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing 100190, China

Abstract. Panel count data occur in many clinical and observational studies and in some situations the observation process is informative. In this article, we propose a new joint model for the analysis of panel count data with time-dependent covariates and possibly in the presence of informative observation process via two latent variables. For the inference on the proposed model, a class of estimating equations is developed and the resulting estimators are shown to be consistent and asymptotically normal. In addition, a lack-of-fit test is provided for assessing the adequacy of the model. The finite-sample behavior of the proposed methods is examined through Monte Carlo simulation studies which suggest that the proposed approach works well for practical situations. Also an illustrative example is provided.
Keywords. Estimating equation; Informative observation process; Joint modeling; Model checking; Panel count data.

1 Introduction

Panel count data usually occur in longitudinal follow-up studies on recurrent events in which study subjects can be observed only at discrete time points rather than continuously. Such data frequently occur in medical periodic follow-up studies, reliability experiments, AIDS

[2]Corresponding author. zhx_math@163.com (H. Zhang)

clinical trials, animal tumorgenicity experiments, and sociological studies (Kalbfleisch and Lawless, 1985; Thall and Lachin, 1988). For the analysis of panel count data, most previous research has been done under the assumption that the recurrent event process and the observation process are completely independent or given covariates. For example, Sun and Kalbfleisch (1995) and Wellner and Zhang (2000) studied nonparametric estimation of the mean function of the underlying counting process arising from panel counts; Sun and Fang (2003), Zhang (2006) and Balakrishnan and Zhao (2009) presented nonparametric tests for the problem of nonparametric comparison of the mean function of counting processes with panel count data. Sun and Wei (2000) and Zhang (2002) investigated regression analysis of panel count data, the former developed some estimating equation-based methods, while the latter proposed a pseudolikelihood approach. Tong et al. (2009) and Zhang, et al. (2013a) studied the variable selection and estimation of panel count data. He et al. (2008) investigated regression analysis of multivariate panel count data.

In many applications, the independence assumption between the recurrent event process and the observation process may not be true. For example, in the bladder cancer study, the occurrence of bladder tumors of a patient and the clinical visit times may be related as discussed in He et al. (2009), Huang et al. (2006), Wellner and Zhang (2007), Liang et al. (2009), Lu et al. (2009) and Sun et al. (2007b), among others. It is well known that when the assumption of non-informative observation times is violated, the methods relying on such an assumption may yield biased results. However, there exists limited research on the analysis of panel count data for the situations where the recurrent event process may be correlated with the observation process given covariates, that is, the observation times may be informative. For example, Huang et al. (2006) studied nonparametric and semiparametric models that allow observation times to be correlated with the recurrent event process through a frailty variable. Sun et al. (2007) investigated semiparametric models for the observation process and the recurrent event process, where both processes may be correlated through a latent variable or frailty. He et al. (2009) proposed some shared frailty models and developed the estimating equations for estimation of regression parameters. More results on related topics can be found on the monograph Sun and Zhao (2013).

Recently, Buzkova (2010) considered panel count data regression with the observation time process are predicted by time-varying factors such as the outcome observed at the last visit. Zhao, et al. (2013) proposed a general and robust estimation approach for regression analysis of panel count data with related observation times. Zhang et al. (2013b) studied a joint model for multivariate panel count data with informative observation times. Here we note that all the proposed method cannot deal with the situation where recurrent event process and informative observation times depend on time-dependent covariates. Thus, there is clearly a need for an analysis method that can directly model panel count data with informative observation times and time-dependent covariates. In this paper, we consider situations in which both the recurrent event data and the observation times depend on time-dependent covariates and propose a new joint model using two latent variables. The proposed joint models are flexible because no parametric assumptions on the distributions of the latent variables are required and the dependence structure of two latent variables is left completely unspecified.

The remainder of the paper is organized as follows. In Section 2, we introduce some notation and describe the proposed models that will be used throughout the paper. In Section 3, an estimating equation approach is developed for the estimation of the regression parameters. Also the asymptotic properties of resulting estimators are established. Section 4 presents a model checking procedure. Section 5 reports some simulation results for assessing the finite sample properties of the proposed estimates. In Section 6, we illustrate our method by applying it to the bladder tumor study. We conclude this article with a discussion section.

2 Statistical Models

Consider a recurrent event study, let $N_i(t)$ denotes the number of occurrences event of interest before or at time t for subject i, $i = 1, \cdots, n$, $0 \le t \le \tau$, where τ is a known constant representing study length. Suppose that for each subject, there exists a $d \times 1$ vector of time-varying covariates denoted by $X_i(t)$. Given $X_i(t)$, two unobserved positive random variables Z_{1i}, Z_{2i}, the mean function of $N_i(t)$ has the form

$$E\{N_i(t)|X_i(t), Z_{1i}, Z_{2i}\} = \mu_0(t)h(Z_{1i})\exp\{\beta' X_i(t)\}, \tag{1}$$

which means that $N_i(t)$ is related to (Z_{1i}, Z_{2i}) through Z_{1i}. Here $\mu_0(\cdot)$ is a completely unknown continuous baseline mean function, $h(\cdot)$ is a completely unspecified positive link function and β is a $d \times 1$ vector of unknown regression parameters. For the identifiability of the model (1), we assume that $E\{h(Z_{1i})|X_i(t)\}$ is a fixed constant, $i = 1, \cdots, n$.

For subject i, we suppose that $N_i(\cdot)$ is observed only at finite time points $T_{i1} < \cdots < T_{iK_i}$, where K_i denotes the potential number of observation times, $i = 1, \cdots, n$. Let C_i be the censoring time, define $\tilde{H}_i(t) = H_i\{\min(t, C_i)\}$, where $H_i(t) = \sum_{j=1}^{K_i} I(T_{ij} \le t)$, $i = 1, \cdots, n$. Then $\tilde{H}_i(t)$ is a point process characterizing the ith subject's observation process and jumps only at the observation times. It is easy to see that $\Delta_i(t)dH_i(t) = d\tilde{H}_i(t)$, where $\Delta_i(t) = I(C_i \ge t)$.

For the observation process, we assume that given $X_i(t)$ and $(Z_{1i},\ Z_{2i})$, $H_i(t)$ follows the following marginal model:

$$E\{dH_i(t)|X_i(t), Z_{1i}, Z_{2i}\} = Z_{2i}\exp\{\gamma' X_i(t)\}d\Lambda_0(t), \tag{2}$$

which means that $H_i(t)$ is related to (Z_{1i}, Z_{2i}) through Z_{2i}. Here $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ is a completely unknown non-decreasing function, γ denotes the vector of $d \times 1$ regression parameters. For identifiability reasons of model (2), we assume that $E(Z_{2i}|X_i(t)) = 1$, $i = 1, \cdots, n$. In the following, it is assumed that $E\{h(Z_1)Z_2|X(t)\} = E\{h(Z_1)Z_2\}$, the censoring time C is independent of (Z_1, Z_2) given $X(t)$, and conditional on $\{X(\cdot), Z_1, Z_2\}$. $N(\cdot)$, $H(\cdot)$ and C are mutually independent.

3 Inference Procedures

In this section, we will consider the estimation procedure for the regression parameters β and γ. Let β_0 and γ_0 denote the true value of β and γ, respectively. Define

$$\mathcal{A}_0(t) = \int_0^t E\{h(Z_1)Z_2\}\mu_0(u)d\Lambda_0(u),$$

and

$$M_i(t;\beta,\gamma,\mathcal{A}) = \int_0^t [N_i(u)\Delta_i(u)dH_i(u) - \Delta_i(u)\exp\{(\beta+\gamma)'X_i(t)\}d\mathcal{A}(u)]. \quad (3)$$

Note that under models (1) and (2), we have

$$\begin{aligned} E\{dM_i(t;\beta_0,\gamma_0,\mathcal{A}_0)|X_i(t),Z_{1i},Z_{2i},C_i\} &= \Delta_i(t)h(Z_{1i})Z_{2i}\exp\{(\beta_0+\gamma_0)'X_i(t)\}\mu_0(t)d\Lambda_0(t) \\ &\quad -\Delta_i(t)\exp\{(\beta_0+\gamma_0)'X_i(t)\}d\mathcal{A}_0(t). \end{aligned}$$

Then, since $E\{h(Z_1)Z_2|X(t)\} = E\{h(Z_1)Z_2\}$ and C is independent of (Z_1, Z_2) given $X(t)$, we get $E\{M_i(t;\beta_0,\gamma_0,\mathcal{A}_0)\} = 0$. That is, the $M_i(t;\beta_0,\gamma_0,\mathcal{A}_0)$'s are zero-mean stochastic process. Thus, for given β and γ, a reasonable estimator for $\mathcal{A}_0(t)$ is the solution to

$$\sum_{i=1}^n M_i(t;\beta,\gamma,\mathcal{A}) = 0, \quad 0 \le t \le \tau.$$

Denote this estimator by $\hat{\mathcal{A}}(t;\beta,\gamma)$, which can be expressed by

$$\hat{\mathcal{A}}(t;\beta,\gamma) = \sum_{i=1}^n \int_0^t \frac{N_i(u)\Delta_i(u)dH_i(u)}{\sum_{j=1}^n \Delta_j(u)\exp\{(\beta+\gamma)'X_j(u)\}}.$$

In view of (3), for a given γ, to estimate β_0, by applying the generalized estimating approach (Liang and Zeger, 1986), we propose the following estimating function for β_0,

$$\begin{aligned} U(\beta;\gamma) &= \frac{1}{n}\sum_{i=1}^n \int_0^\tau W(t)\{X_i(t) - \bar{X}(t;\beta,\gamma)\}[N_i(t)\Delta_i(t)dH_i(t) \\ &\quad - \exp\{(\beta+\gamma)'X_i(t)\}\Delta_i(t)d\hat{\mathcal{A}}(t;\beta,\gamma)\}], \quad (4) \end{aligned}$$

where $W(t)$ is a possibly data-dependent weight function,

$$\bar{X}(t;\beta,\gamma) = \frac{\sum_{i=1}^{n} \Delta_i(t) X_i(t) \exp\{(\beta+\gamma)' X_i(t)\}}{\sum_{j=1}^{n} \Delta_j(t) \exp\{(\beta+\gamma)' X_j(t)\}}.$$

Of course, in reality, γ is unknown. It is easy to see that model (2) implies

$$E\{dH_i(t)|X_i(t)\} = \exp\{\gamma' X_i(t)\} d\Lambda_0(t). \tag{5}$$

Then by following the approach proposed by Lin et al. (2000) for the proportional rate model, γ can be consistently estimated by the following estimating equation:

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{X_i(t) - \bar{X}^*(t;\gamma)\}\Delta_i(t) dH_i(t) = 0, \tag{6}$$

where

$$\bar{X}^*(t;\gamma) = \frac{\sum_{i=1}^{n} \Delta_i(t) \exp\{\gamma' X_i(t)\} X_i(t)}{\sum_{j=1}^{n} \Delta_j(t) \exp\{\gamma' X_j(t)\}}.$$

Let $\hat{\gamma}$ denote the solution to the above estimating equation (6). Given $\hat{\gamma}$, we can estimate β_0 by solving the equation $U(\beta;\hat{\gamma}) = 0$. The asymptotic normality of $\hat{\beta}$ is established in the following theorem, with the proof given in the Appendix.

Theorem 1. *Under the regularity conditions (C1)-(C4) stated in the Appendix, then as $n \to \infty$, we have*

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{D}} N(0, A^{-1}\Sigma A^{-1}),$$

where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution, A and Σ are defined in the Appendix.

Define

$$\begin{aligned}
S^{(k)}(t;\gamma) &= \frac{1}{n}\sum_{i=1}^{n} \Delta_i(t) \exp\{\gamma' X_i(t)\} X_i(t)^{\otimes k}, k = 0, 1, 2, \\
\hat{\Lambda}_0(t) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{t} \frac{\Delta_i(u) dH_i(u)}{S^{(0)}(u;\hat{\gamma})}, \\
\hat{\mathcal{M}}_i(t) &= H_i(t \wedge C_i) - \int_0^{t} \Delta_i(u) \exp\{\hat{\gamma}' X_i(u)\} d\hat{\Lambda}_0(u),
\end{aligned}$$

where $a \wedge b = \min(a, b)$. Then the covariance matrix in the above theorem can be consistently estimated by $\hat{A}^{-1}\hat{\Sigma}\hat{A}^{-1}$, where $\hat{\Sigma} = n^{-1}\sum_{i=1}^{n}\hat{\xi}_i^{\otimes 2}$,

$$\begin{aligned}
\hat{\xi}_i &= \int_0^\tau W(t)\{X_i(t) - \bar{X}(t;\hat{\beta},\hat{\gamma})\}[N_i(t)\Delta_i(t)dH_i(t) - \exp\{(\hat{\beta}+\hat{\gamma})'X_i(t)\}\Delta_i(t)d\hat{\mathcal{A}}(t)] - \hat{A}\hat{Q}^{-1}\hat{\eta}_i, \\
\hat{\eta}_i &= \int_0^\tau \{X_i(t) - \bar{X}(t;\hat{\gamma})\}d\hat{\mathcal{M}}_i(t), \\
\hat{A} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau W(t)\{X_i(t) - \bar{X}(t;\hat{\beta},\hat{\gamma})\}^{\otimes 2}\Delta_i(t)\exp\{(\hat{\beta}+\hat{\gamma})'X_i(t)\}d\hat{\mathcal{A}}(t),
\end{aligned}$$

and

$$\hat{Q} = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left[\frac{S^{(2)}(t;\hat{\gamma})}{S^{(0)}(t;\hat{\gamma})} - \bar{X}^*(t;\hat{\gamma})^{\otimes 2}\right]\Delta_i(t)dH_i(t).$$

4 Model checking

In this section, we will consider the checking of the adequacy of models (1) and (2). Motivated by Lin et al. (2000), we consider the statistic:

$$\Phi(t) = n^{-1/2}\sum_{i=1}^{n}\int_0^t g(X_i(u))d\hat{M}_i(u),$$

where g is a known bounded function, $\hat{M}_i(t) = M_i(t;\hat{\beta},\hat{\gamma},\hat{\mathcal{A}}(\cdot;\hat{\beta},\hat{\gamma}))$. To derive the distribution of $\Phi(t)$, define

$$\begin{aligned}
S(u;\hat{\beta},\hat{\gamma}) &= \frac{1}{n}\sum_{j=1}^{n} g(X_j(u))\Delta_j(u)\exp\{(\hat{\beta}+\hat{\gamma})'X_j(u)\}, \\
S_0(u;\hat{\beta},\hat{\gamma}) &= \frac{1}{n}\sum_{j=1}^{n}\Delta_j(u)\exp\{(\hat{\beta}+\hat{\gamma})'X_j(u)\},
\end{aligned}$$

and

$$B(t;\hat{\beta},\hat{\gamma}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \left[g(X_i(u)) - \frac{S(u;\hat{\beta},\hat{\gamma})}{S_0(u;\hat{\beta},\hat{\gamma})}\right]\Delta_i(u)X_i(u)\exp\{(\hat{\beta}+\hat{\gamma})'X_i(u)\}d\hat{\mathcal{A}}(u).$$

We will show in the Appendix that the null distribution of $\Phi(t)$ can be approximated by the zero-mean Gaussian process

$$\begin{aligned}\tilde{\Phi}(t) &= n^{-1/2}\sum_{i=1}^{n}\int_0^t\left[g(X_i(u))-\frac{S(u;\hat{\beta},\hat{\gamma})}{S_0(u;\hat{\beta},\hat{\gamma})}\right]d\hat{M}_i(u)-B(t;\hat{\beta},\hat{\gamma})'\hat{A}^{-1}n^{-1/2}\sum_{i=1}^{n}\hat{\xi}_i \\ &\quad -B(t;\hat{\beta},\hat{\gamma})'\hat{Q}^{-1}n^{-1/2}\sum_{i=1}^{n}\hat{\eta}_i+o_p(1). \qquad (7)\end{aligned}$$

Following the resampling approach presented in Lin et al. (2000), let $(G_1,\cdots,G_n)$ be independent standard normal variables independent of the data. Then we can show that the null distribution of $\Phi(t)$ can be approximated by the zero-mean Gaussian process,

$$\begin{aligned}\hat{\Phi}(t) &= n^{-1/2}\sum_{i=1}^{n}\int_0^t\left[g(X_i(u))-\frac{S(u;\hat{\beta},\hat{\gamma})}{S_0(u;\hat{\beta},\hat{\gamma})}\right]d\hat{M}_i(u)G_i-B(t;\hat{\beta},\hat{\gamma})'\hat{A}^{-1}n^{-1/2}\sum_{i=1}^{n}\hat{\xi}_iG_i \\ &\quad -B(t;\hat{\beta},\hat{\gamma})'\hat{Q}^{-1}n^{-1/2}\sum_{i=1}^{n}\hat{\eta}_iG_i.\end{aligned}$$

We note that $\Phi(t)$ is expected to fluctuate randomly around 0 under model (1), thus we can construct a lack-of-fit test based on the supremum statistic $\sup_{0\le t\le\tau}|\Phi(t)|$, where the p-value can be obtained by comparing the observed value of $\sup_{0\le t\le\tau}|\Phi(t)|$ to a large number of realizations from $\sup_{0\le t\le\tau}|\hat{\Phi}(t)|$.

5 A simulation study

In this section, an extensive simulation studies was conducted to evaluate the performance of the proposed inference procedure in Section 4 with the focus on estimation of β. In the study, we assumed the time-dependent covariate $X_i(t)$ was taken as $X_i\log(t)$ with X_i generated form a uniform distribution on $(0,1)$. The censoring time C_i was taken as $\min(C_i^*,\tau)$ with C_i^* following the uniform distribution over (10, 20) and $\tau=15$. We generated the latent variable Z_{2i} from a uniform distribution on (0.5, 1.5). For given Z_{2i}, we generated the latent variable Z_{1i} by $Z_{1i}=Z_{2i}^{\alpha}$ with $\alpha=0.5$, 0 and -0.5. Here, when $\alpha>0$, a subject with more frequent observations would have a higher occurrence rate of the recurrent event and the two processes are positively correlated; when $\alpha = 0$, the two processes have no correlation given the covariates; when $\alpha < 0$, a subject with more frequent observations would have a lower occurrence rate of the recurrent event and the two processes are negatively correlated.

For the observation process, we assumed that the observation times H_i follows a non-homogeneous Poisson process with the marginal (2), where $\gamma_0 = 1$ and $\Lambda_0(t) = \frac{t}{10}$. Let $m_i(t) = Z_{2i} \exp\{\gamma_0' X_i(t)\}\Lambda_0(t)$, then given Z_{2i}, $X_i(t)$ and C_i, the observation times $\{T_{i,j}\}$ for the ith subject over $[0, C_i]$ can be generated by the thinning algorithm (Ross, 2006):

Step 1: $t = 0$, $j = 0$ and $T_{i,j} = 0$.

Step 2: Generate a random number u from uniform (0,1) distribution.

Step 3: $t = t - \frac{1}{m_i(C_i)} \log(u)$.

Step 4: Generate a random number v from uniform (0,1) distribution.

Step 5: If $v \leq \frac{m_i(t)}{m_i(C_i)}$, set $j = j + 1$ and $T_{i,j} = t$.

Step 6: Go to Step 2.

To generate $N_i(T_{ik})$'s, for given $T_{i,j}$, we assume that

$$N_i(T_{i,j}) = N_i(T_{i,1}) + \{N_i(T_{i,2}) - N_i(T_{i,1})\} + \cdots + \{N_i(T_{i,j}) - N_i(T_{i,j-1})\},$$

where $T_{i,0} = 0$, $N_i(t) - N_i(s)$ follows the Poisson distribution with mean

$$Z_{1i}[\mu_0(t) \exp\{\beta X_i(t)\} - \mu_0(s) \exp\{\beta X_i(s)\}]$$

with $\mu_0(t) = t$. All the results reported here are based on 500 Monte Carlo replications using MATLAB software.

In Table 1, we presented the simulation results obtained on the estimation of β with the sample size $n = 100$ or 200. It includes the estimated bias (BIAS) given by the average of the proposed estimates of minus the true value, the sample standard error (SSE) of the proposed estimates, the mean of the estimated standard error (ESE), and the empirical 95% coverage probabilities (CP). It can be seen from the table that the point estimates seem to be unbiased and that SEE and SSE are quite close to each other, indicating that the proposed variance estimation seems to work well. Figure 1 reports the quantile plot for the estimate for β with $\alpha = 0.5$, $\beta = 1$, $n = 200$. Similar quantile plots were obtained for other setups. Also the results indicate that the normal approximation seems to be appropriate and as expected, the results became better overall as the sample size n increased.

6 An illustrative example

In this section, we apply the proposed methods to the bladder cancer data, which have been discussed earlier and analysed by Sun et al. (2005), Huang et al. (2006), Sun et al. (2007) and Liang et al. (2009), among others. The data set includes 85 bladder cancer patients, 47 in the placebo group and 38 in the thiotepa treatment group. For each patient, the observed information includes the clinical visit or observation times in a month and the number of bladder tumours that occurred between clinical visits. The longest observation time is 53 months. In addition, two baseline covariates were measured, namely, the number of initial tumours that the patients had before entering the study and the size of the largest initial tumour. One of the main objectives of the study was to assess the effect of the thiotepa treatment on bladder tumour recurrence. The size of the largest initial tumour has been shown to have no effect on the recurrence rate (Sun and Wei, 2000; Zhang, 2002). Therefore, in the current study, we focus on the effects of thiotepa treatment and the number of initial tumours on the recurrence rate of bladder tumour while allowing for a possible correlation between the tumour recurrence and the clinical visit process.

For the analysis, we define $X_i = (X_{i1}, X_{i2})'$ with $X_{i1} = 1$ if subject i was in the thiotepa treatment group and 0 otherwise and X_{i2} denoting the number of initial tumors of the ith patient. The application of the estimation procedure proposed in the previous sections gave $\hat{\beta} = (-1.4496, 0.2399)'$ with the estimated standard errors of 0.3538 and 0.0614, respectively. They indicate that the thiotepa treatment had a significant effect in reducing the occurrence rate of the bladder tumor and the occurrence rate was significantly positively related to the number of initial tumors. It can be seen that the conclusion obtained here is similar to Zhao, et al. (2013).

To check the goodness-of-fit of the models (1) and (2), we used the omnibus test procedure given in Section 4. Since the two covariates are time-invariant, we used the following supremum test statistic $\sup_{0\le t\le\tau, x_1, x_2} |\Phi(t, x_1, x_2)|$, where

$$\Phi(t, x_1, x_2) = n^{-1/2} \sum_{i=1}^{n} I(X_{i1} \le x_1, X_{i2} \le x_2)\hat{M}_i(t),$$

where the event $I(X_{i1} \le x_1, X_{i2} \le x_2)$ means that each component of X_{i1} and X_{i2} is no larger than the corresponding component of x_1 and x_2. We obtained the p-value of 0.368 based on 1000 realizations. This suggests that these models seem to be appropriate for the bladder cancer data considered here.

7 Concluding remarks

In this paper, we have considered regression analysis of panel count data with time-dependent covariates and possibly in the presence of informative observation process via two latent variables. A key advantage of the proposed approach over existing methods for panel count data is that the proposed joint models are flexible because no parametric assumptions on the distributions of the latent variables are required and the dependence structure of two latent variables is left completely unspecified. For estimation of regression parameters, we have developed an estimating equation approach that yields consistent and asymptotically normal parameter estimates. The simulation results suggest that the proposed inference procedures perform well and an illustrative example is also presented.

There exist several directions for future research. In the proposed methodology, we assumed that the censoring times C_i are independent of covariates. However, this assumption may be relaxed. To generalize the proposed methodology to the situation where C_i may depend on covariates, one approach is to specify a regression model such as the proportional hazards model. Another interesting topic is that we can generalized the proposed procedure to the time-varying coefficient model,

$$E\{N_i(t)|X_i(t), Z_{1i}(t), Z_{2i}(t)\} = \mu_0(t)h\{Z_{1i}(t)\}\exp\{\beta(t)'X_i(t)\}, \tag{8}$$

where $\beta(t)$, $Z_{1i}(t)$ and $Z_{2i}(t)$ are defined as before except being time-dependent. Moreover, we may study the inference for the above (8) with terminal event(Sun, et al. 2012).

Acknowledgments

The authors would like to thank a referee for the valuable comments, which greatly improved the paper. The work of Haixiang Zhang is partially supported by the National Natural Science Foundation of China (No. 11301212) and China Postdoctoral Science Foundation (No. 2014M550861). The work of Liuquan Sun is partially supported by the National Natural Science Foundation of China Grants (No. 11231010, 11171330 and 11021161) and Key Laboratory of RCSDS, CAS (No.2008DP173182). The work of Dehui Wang is partly supported by National Natural Science Foundation of China (11271155), Specialized Research Fund for the Doctoral Program of Higher Education (20110061110003), Scientific Research Fund of Jilin University (201100011) and Jilin Province Natural Science Foundation (20101596).

Appendix

In this section, we need the following regularity conditions:

(C1) $P(C \geq \tau) > 0$, and $H(\tau)$ is bounded by a constant.

(C2) $X(t)$ has bounded variation on $[0, \tau]$.

(C3) The weight function $W(t)$ has bounded variation and converges to a deterministic function $w(t)$ in probability uniformly in $t \in [0, \tau]$.

(C4) A is nonsingular, where

$$A = E\left[\int_0^\tau \{X_i(t) - \bar{x}(t)\}^{\otimes 2}\Delta_i(t)\exp\{(\beta_0 + \gamma_0)'X_i(t)\}d\mathcal{A}_0(t)\right],$$

and $\bar{x}(t)$ is the limit of $\bar{X}(t;\beta_0,\gamma_0)$.

Proof of Theorem 1. Following the arguments similar to those given in Appendix A.2 of Lin and Ying (2001), we have

$$\begin{aligned} n^{1/2}U(\beta_0;\gamma_0) &= n^{-1/2}\sum_{i=1}^{n}\int_0^{\tau} w(t)\{X_i(t)-\bar{x}(t)\}[N_i(t)\Delta_i(t)dH_i(t) \\ &\quad -\exp\{(\beta_0+\gamma_0)'X_i(t)\}\triangle_i(t)d\mathcal{A}_0(t)]+o_p(1), \end{aligned} \tag{9}$$

which is a sum of n independent zero-mean random vectors plus an asymptotically negligible term.

It is easy to see that $-\partial\hat{U}(\beta_0;\gamma)/\gamma|_{\gamma=\gamma_0}$ converges in probability to A. Furthermore, it follows from Lin et al. (2000) that

$$n^{1/2}(\hat{\gamma}-\gamma_0)=Q^{-1}n^{-1/2}\sum_{i=1}^{n}\eta_i+o_p(1), \tag{10}$$

where

$$\eta_i=\int_0^{\tau}\{X_i(t)-\bar{x}^*(t)\}d\mathcal{M}_i(t),$$

$$Q=E\left[\int_0^{\tau}\{X_i(t)-\bar{x}^*(t)\}^{\otimes 2}\Delta_i(t)\exp\{\gamma_0'X_i(t)\}d\Lambda_0(t)\right],$$

$\bar{x}^*(t)$ is the limit of $\bar{X}^*(t;\gamma_0)$, and

$$\mathcal{M}_i(t)=H_i(t\wedge C_i)-\int_0^{t}\Delta_i(u)\exp\{\gamma_0'X_i(u)\}d\Lambda_0(u).$$

Then by the Taylor series expansion of $U(\beta_0;\hat{\gamma})$ at $U(\beta_0;\gamma_0)$ and the consistency of $\hat{\gamma}$, we have

$$n^{1/2}U(\beta_0;\hat{\gamma})=n^{1/2}U(\beta_0;\gamma_0)-An^{1/2}(\hat{\gamma}-\gamma_0)+o_p(1). \tag{11}$$

Let

$$\xi_i=\int_0^{\tau} w(t)\{X_i(t)-\bar{x}(t)\}[N_i(t)\Delta_i(t)dH_i(t)-\exp\{(\beta_0+\gamma_0)'X_i(t)\}\Delta_i(t)d\mathcal{A}_0(t)]-AQ^{-1}\eta_i. \tag{12}$$

Thus, it follows form (9), (11) and (12) that

$$n^{1/2}U(\beta_0;\hat{\gamma}) = n^{-1/2}\sum_{i=1}^{n}\xi_i.$$

By the multivariate central theorem, we know that $n^{1/2}U(\beta_0;\hat{\gamma})$ converges in distribution to a zero-mean normal random vector with covariance matrix $\Sigma = E(\xi_i^{\otimes 2})$. Moreover, we note that $-\partial U(\beta_0;\hat{\gamma})/\partial\beta$ also converges in probability to A. Then by the Taylor series expansion of $\hat{U}(\hat{\beta};\hat{\gamma})$ at $U(\beta_0;\hat{\gamma})$, we have

$$n^{1/2}(\hat{\beta}-\beta_0) = A^{-1}n^{1/2}U(\beta_0;\hat{\gamma}) + o_p(1). \tag{13}$$

Thus, $n^{1/2}(\hat{\beta}-\beta_0)$ is asymptotically zero-mean normal with covariance matrix $A^{-1}\Sigma A^{-1}$. □

Proof of asymptotic properties of $\Phi(t)$. In the following, let $s(t)$, $s_0(t)$, $b(t)$ be the limits of $S(t;\beta_0,\gamma_0)$, $S_0(t;\beta_0,\gamma_0)$, $B(t;\beta_0,\gamma_0)$, respectively. It then follows from Lemma A.1 of Lin and Ying (2001) and the functional version of the Taylor expansion that

$$\Phi(t) = n^{-1/2}\sum_{i=1}^{n}\int_0^t\left[g(X_i(u)) - \frac{s(u)}{s_0(u)}\right]dM_i(t) - b(t)'n^{1/2}(\hat{\beta}-\beta_0) - b(t)'n^{1/2}(\hat{\gamma}-\gamma_0) + o_p(1).$$

Then from (10) and (13) that $\Phi(t)$ can be written as

$$\begin{aligned}\Phi(t) &= n^{-1/2}\sum_{i=1}^{n}\int_0^t\left[g(X_i(u)) - \frac{s(u)}{s_0(u)}\right]dM_i(t) - b(t)'A^{-1}n^{-1/2}\sum_{i=1}^{n}\xi_i \\ &\quad -b(t)'Q^{-1}n^{-1/2}\sum_{i=1}^{n}\eta_i.\end{aligned}$$

Thus, it follows from the multivariate central theorem that $\Phi(t)$ converges to a finite-dimensional zero-mean Gaussian process. Because any bounded variance function can be written as the difference of two increasing functions, the first term of $\Phi(t)$ is tight. Since $b(t)$ is deterministic function and ξ_i and η_i do not involve t, thus the second and third terms of $\Phi(t)$ are tight. Thus, $\Phi(t)$ is tight and converges weakly to a zero-mean Gaussian process that can be approximated by the zero-mean Gaussian process $\tilde{\Phi}(t)$ given by (7). □

References

Balakrishnan, N., Zhao, X. (2009). New multi-sample nonparametric tests for panel count data. *Annals of Statistics*, **37**, 1112-1149.

Buzkova, P., (2010). Panel count data regression with informative observation times. *The International Journal of Biostatistics*, **6** (1). Article 30.

He, X., Tong, X., Sun, J. and Cook, R. (2008). Regression analysis of multivariate panel count data. *Biostatistics*, **9**, 234-248.

He, X., Tong, X., Sun, J. (2009). Semiparametric analysis of panel count data with correlated observation and follow-up times. *Lifetime Data Analysis*, **15**, 177-196.

Huang, C.Y., Wang, M.C., Zhang, Y., (2006). Analysing panel count data with informative observation times. *Biometrika*, **93**, 763-775.

Kalbfleisch, J.D., Lawless, J.F., 1985. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, **80**, 863-871.

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.

Lin, D. Y., Wei, L. J. and Ying, Z. (2001). Semiparametric transformation models for point processes. *Journal of the American Statistical Association*, **96**, 620-628.

Lin, D. Y. and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, **96**, 103-126.

Ross, S. (2006) *Simulation*, 4th edn. Academic Press, New York.

Sun, J., Fang, H.B., (2003). A nonparametric test for panel count data. *Biometrika*, **90**, 199-208.

Sun, J., Kalbfleisch, J.D., (1995). Estimation of the mean function of point processes based on panel count data. *Statistica Sinica*, **5**, 279-290.

Sun, J., Tong, X., He, X., (2007). Regression analysis of panel count data with dependent observation times. *Biometrics*, **63**, 1053-1059.

Sun, J., 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.

Sun, J., Zhao, X., (2013). *The Statistical Analysis of Panel Count Data*. Springer Science + Business Inc.

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

Thall, P.F., Lachin, J.M., 1988. Analysis of recurrent events: nonparametric methods for random-interval count data. *Journal of the American Statistical Association*, **83**, 339-347.

Tong, X., He, X., Sun, L. and Sun, J. (2009).Variable selection for panel count data via non-concave penalized estimating function. *Scandinavian Journal of Statistics*, **36**, 620-635.

Wellner, J.A., Zhang, Y., 2000. Two estimators of the mean of a counting process with panel count data. *Annals of Statistics*, **28**, 779-814

Zhang, H., Sun, J. and Wang, D. (2013a). Variable selection and estimation for multivariate panel count data via the seamless-L_0 penalty. *The Canadian Journal of Statistics*, **41**(2), 368-385.

Zhang, H., Zhao, H., Sun, J., Wang, D. and Kim, K.(2013b). Regression analysis of multivariate panel count data with an informative observation process. *Journal of Multivariate Analysis*, **119**, 71-80.

Zhang, Y., 2006. Nonparametric k-sample tests for panel count data. *Biometrika*, **93**, 777-790.

Zhao, X., Tong, X., 2011. Semiparametric regression analysis of panel count data with information observation times. *Computational Statistics and Data Analysis*, **55**, 291-300.

Zhao, X., Tong, X. and Sun, J. (2013). Robust estimation for panel count data with informative observation times. *Computational Statistics and Data Analysis*, **57**, 33-40.

Table 1. Simulation results on the estimation of β.

			$\alpha = -0.5$			
β	1	0	-1	1	0	-1
		n=100			n=200	
BIAS	0.0049	0.0053	-0.0034	-0.0039	0.0047	-0.0007
SSE	0.1176	0.0793	0.1454	0.0789	0.0543	0.1016
ESE	0.1085	0.0725	0.1379	0.0769	0.0526	0.0988
CP	0.934	0.926	0.932	0.954	0.936	0.940
			$\alpha = 0$			
β	1	0	-1	1	0	-1
		n=100			n=200	
BIAS	-0.0039	0.0017	0.0021	-0.0000	-0.0004	-0.0001
SSE	0.1119	0.0739	0.1422	0.0826	0.0521	0.0961
ESE	0.1092	0.0675	0.1345	0.0797	0.0488	0.0975
CP	0.946	0.930	0.932	0.950	0.932	0.956
			$\alpha = 0.5$			
β	1	0	-1	1	0	-1
		n=100			n=200	
BIAS	-0.0059	-0.0025	-0.0002	-0.0065	-0.0051	-0.0002
SSE	0.1378	0.0771	0.1395	0.0923	0.0528	0.0985
ESE	0.1221	0.0709	0.1396	0.0894	0.0514	0.0976
CP	0.920	0.928	0.938	0.952	0.948	0.950

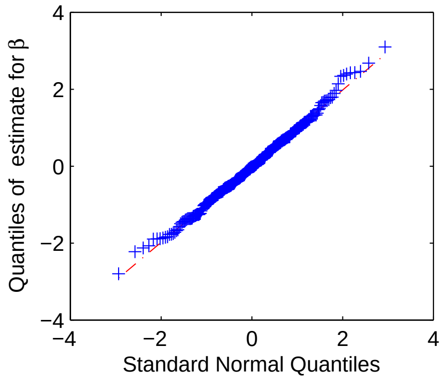

Figure 1. Q-Q plot of standardized estimates for β, where $\alpha = 0.5$, $\beta = 1$, $n = 200$.