Computational Statistics and Data Analysis 57 (2013) 33–40

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Robust estimation for panel count data with informative observation times

Xingqiu Zhao [a,*], Xingwei Tong [b], Jianguo Sun [c]

[a] Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong
[b] Department of Statistics, Beijing Normal University, Beijing, 100875, China
[c] Department of Statistics, University of Missouri, Columbia, MO, 65211, USA

ARTICLE INFO

Article history:
Received 22 September 2011
Received in revised form 14 May 2012
Accepted 17 May 2012
Available online 30 May 2012

Keywords:
Informative observation process
Panel count data
Recurrent event process
Robust estimation

ABSTRACT

Panel count data usually occur in longitudinal follow-up studies that concern occurrence rates of certain recurrent events and their analysis involves two processes. One is the underlying recurrent event process of interest and the other is the observation process that controls observation times. In some situations, the two processes may be correlated and, for this, several estimation procedures have recently been developed (He et al., 2009; Huang et al., 2006; Sun et al., 2007b; Zhao and Tong, 2011). These methods, however, rely on some restrictive models or assumptions such as the Poisson assumption. In this work, a more general and robust estimation approach is proposed for regression analysis of panel count data with related observation times. The asymptotic properties of the resulting estimates are established and the numerical studies conducted indicate that the approach works well for practical situations.

1. Introduction

This work considers regression analysis of panel count data when the observation times or process may be related to the underlying recurrent event process of interest governing the panel count data. By panel count data, we mean the data that concern occurrence rates of certain recurrent events and give only the numbers of the events that occur between the observation times, but not their occurrence times. Such data naturally occur in longitudinal follow-up studies on recurrent events in which study subjects can be observed only at discrete time points rather than continuously (Cai and Schaubel, 2004; Cook and Lawless, 2007; Sun, 2006).

Many authors have discussed the analysis of panel count data when the recurrent event process of interest and the observation process are independent completely or conditional on covariates. In this case, the inference can be made conditional on the observation process. For example, Sun and Kalbfleisch (1995), Wellner and Zhang (2000) and Hu et al. (2009) studied nonparametric estimation of the mean function of the underlying counting process yielding panel counts. The same problem was also considered by Zhang and Jamshidian (2003) and Lu et al. (2007). The former employed a gamma frailty variable to account for the correlation among panel counts and developed a maximum pseudo-likelihood approach, while the latter also gave some likelihood-based estimators of the mean functions by using monotone polynomial splines. In addition, Sun and Fang (2003), Zhang (2006) and Balakrishnan and Zhao (2009) constructed some nonparametric tests for nonparametric comparison of the mean functions of counting processes. Cheng and Wei (2000), Sun and Wei (2000) and Hu et al. (2003) proposed some semiparametric models for regression analysis of panel count data and developed some

* Corresponding author. Tel.: +852 27666921; fax: +852 23629045.
E-mail addresses: xingqiu.zhao@polyu.edu.hk (X. Zhao), xweitong@bnu.edu.cn (X. Tong), sunj@missouri.edu (J. Sun).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.05.015

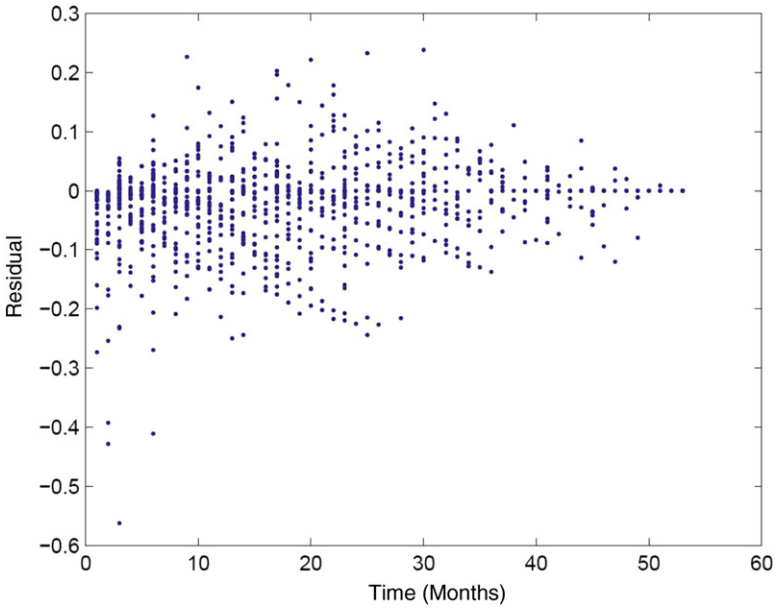

Fig. 1. Plot of the residuals for fitted model (2) with the bladder tumor data.

estimating equation-based approaches. Zhang (2002), Wellner and Zhang (2007) and Lu et al. (2009) discussed the same regression problem and gave some semiparametric likelihood-based approaches.

Sometimes the recurrent event process of interest and the observation process may be related. A well-known example of such panel count data is the bladder cancer data 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. The data concern the occurrence rate of bladder tumors and during the study giving rise to the data, study patients were observed periodically and at different time points. Some patients were observed more often than others and several authors have showed that the occurrence process of bladder tumors seems to be correlated with the observation process. More details about the example are given in Section 5. For the analysis of such panel count data, several methods have been developed (He et al., 2009; Huang et al., 2006; Sun et al., 2007b; Zhao and Tong, 2011). A common and key assumption behind these methods is that the observation process is a Poisson process.

To be more specific about this, consider a recurrent event study and suppose that only panel count data are available. Let $N(t)$ and $O(t)$ denote the process of interest and the observation process, respectively, and X a vector of the covariates of interest. Then $N(t)$ is observed only at the times where $O(t)$ jumps. Suppose that $N(t)$ and $O(t)$ may be related even given X. For inference, Sun et al. (2007b) assumed that given X and a latent variable Z, the mean function of $N(t)$ has the form

$$E\{N(t)|X, Z\} = Z^{\alpha}\, \mu_0(t) \exp(X'\beta) \tag{1}$$

and $O(t)$ is a non-homogeneous Poisson process with the intensity function

$$\lambda(t|X, Z) = Z\, \lambda_0(t) \exp(X'\gamma). \tag{2}$$

In the above, β, α and γ are unknown parameters, and $\mu_0(t)$ and $\lambda_0(t)$ are unknown baseline mean and intensity functions, respectively. To examine the Poisson process assumption for the bladder cancer data, we fitted the data on the observation times to model (2) and present the residuals in Fig. 1. Also we developed a simple Kolomogorov–Smirnov test statistic procedure (Gibbons and Chakraborti, 2011) and obtained the p-value of 0.07 for testing the Poisson process assumption. Both the figure and the test suggest that the Poisson process assumption with model (2) may be questionable. In the following, we will relax this and other assumptions and develop a general and robust inference approach.

Note that a number of methods have been developed for regression analysis of longitudinal data, mostly under the assumption that the longitudinal response and the observation process are independent completely or given covariates. For example, Diggle et al. (1994) provided a comprehensive summary about the commonly used methods such as estimating equation and random effect model approaches, and Lin and Ying (2001) and Welsh et al. (2002) discussed general semiparametric regression analysis of longitudinal data. In contrast, limited research exists for regression analysis of longitudinal data where measurement times may be informative or still related to the underlying longitudinal process even given covariates, a problem similar to that discussed here (Sun et al., 2007a; Liu et al., 2008; Liang et al., 2009). Although panel count data can be regarded as a special type of longitudinal data, the use of the methods developed for longitudinal data may not be valid or efficient as they do not take into account the special structure of panel count data.

The remainder of this work is organized as follows. In Section 2, we will begin with introducing the notation and assumptions and then presenting the models that will be used below. The models include models (1) and (2) above as special cases. In particular, for the observation process, we employ a rate function model instead of model (2) and do not require $O(t)$ to be a Poisson process. A robust estimation procedure is presented in Section 3 for the parameters of interest and the asymptotic properties of the resulting estimators are established. Also a model check procedure is presented. Section 4 reports some simulation results obtained for assessing the finite sample properties of the proposed estimates and an illustrative example is given in Section 5.

2. Notation, assumptions and models

Consider a recurrent event study that consists of n independent subjects and let $N_i(t)$ denote the number of occurrences of the recurrent event of interest before or at time t for subject i. Suppose that for each subject, there exists a p-dimensional vector of covariates denoted by X_i. Given X_i and an unobserved positive random variable Z_i that is independent of X_i, the mean function of $N_i(t)$ has the form

$$E\{N_i(t)|X_i, Z_i\} = \mu_0(t)\, g(Z_i) \exp(X_i'\beta). \tag{3}$$

Zhao and Tong (2011). Here as in model (1), $\mu_0(t)$ is a completely unknown continuous baseline mean function, β is a vector of unknown regression parameters, and $g(\cdot)$ is a completely unspecified link function. It is easy to see that this model includes model (1) as a special case.

For subject i, 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, $i = 1, \ldots, n$. That is, only the values of $N_i(t)$ at these observation times are known and we have panel count data on the $N_i(t)$'s. Let C_i denote the follow-up time associated with subject i and thus $N_i(t)$ is observed only at these T_{ij}'s with $T_{ij} \le C_i$, $i = 1, \ldots, n$. Define $\tilde{O}_i(t) = O_i(\min(t, C_i))$, where $O_i(t) = \sum_{j=1}^{K_i} I(T_{ij} \le t)$, $i = 1, \ldots, n$. Then $\tilde{O}_i(t)$ is a point process characterizing the ith subject's observation process and jumps only at the observation times.

For the observation process, instead of model (2), we will assume that $O_i(t)$ satisfies the following rate function model

$$E\{dO_i(t)|X_i, Z_i\} = Z_i\, h(X_i)\, d\Lambda_0(t), \tag{4}$$

where $h(\cdot)$ is a completely unspecified positive function as g and $\Lambda_0(\cdot)$ is a completely unknown continuous baseline function. It is easy to see that model (2) implies

$$E\{dO_i(t)|O_i(s), 0 \le s < t, X_i, Z_i\} = E\{dO_i(t)|X_i, Z_i\}$$

corresponding to the independent increment structure of the Poisson process. Under model (4), one does not need this assumption anymore. In the following, it will be assumed that given (X_i, Z_i), $N_i(t)$ and $O_i(t)$ are independent. Also C_i is independent of $\{N_i, O_i, X_i, Z_i\}$ and $\{N_i(t), O_i(t), C_i, X_i, 0 \le t \le \tau\}_{i=1}^n$ are independent and identically distributed, where τ denotes the length of the study. Suppose that the main goal is to estimate regression parameter β.

3. Inference procedure

To estimate β, note that if the latent variables Z_i's are known, model (3) would become the usual proportional means model and several methods such as that given in Cheng and Wei (2000) can be used. Unfortunately, the Z_i's are unknown in practice. One natural way for this is to estimate the Z_i's first and then treat them as known. In the following, we take a different approach motivated by that proposed in Sun and Wei (2000) among others.

Specifically, define

$$\bar{N}_i = \sum_{j=1}^{m_i} N_i(T_{ij}) I(T_{ij} \le C_i) = \int_0^\tau N_i(t) d\tilde{O}_i(t),$$

where $m_i = \tilde{O}_i(C_i)$, the total number of observations on subject i, $i = 1, \ldots, n$. Then, we have

$$E\left(\bar{N}_i|X_i\right) = \exp(\beta' X_i) h(X_i) E\{g(Z_i)Z_i\} \int_0^\tau P(C_i \ge t)\mu_0(t) d\Lambda_0(t)$$

and

$$E(m_i|X_i) = E(Z_i)E\{\Lambda_0(C_i)\}h(X_i).$$

These yield

$$E(\bar{N}_i|X_i) = E(m_i|X_i) \exp(\beta' X_i + \theta),$$

where

$$\theta = \log\left[\frac{E\{g(Z_i)Z_i\}}{E(Z_i)E\{\Lambda_0(C_i)\}} \int_0^\tau P(C_i \ge t)\mu_0(t) d\Lambda_0(t)\right],$$

an unknown parameter. For estimation of β, motivated by the equation above, we propose to use the following class of estimating functions

$$U(\beta_1) = \sum_{i=1}^{n} W_i X_{1i}\{\bar{N}_i - m_i \exp(\beta_1' X_{1i})\} = 0, \tag{5}$$

where the W_i's are some weights that could depend on X_i, $X_{1i}' = (X_i', 1)$ and $\beta_1' = (\beta', \theta)$.

Let $\hat{\beta}_1 = (\hat{\beta}', \hat{\theta})'$ denote the solution to Eq. (5) and $\beta_{10} = (\beta_0', \theta_0)'$ the true value of β_1. Then, we will show in the Appendix A.1 that under some regularity conditions, the estimator $\hat{\beta}_1$ is consistent and $\sqrt{n}(\hat{\beta}_1 - \beta_{10})$ has asymptotically a normal distribution with mean zero and the covariance matrix that can be consistently estimated by $\hat{\Gamma}^{-1}\hat{\Sigma}\hat{\Gamma}^{-1}$, where

$$\hat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n}\left\{W_i m_i X_{1i} X_{1i}' \exp(\hat{\beta}_1' X_{1i})\right\}$$

and $\hat{\Sigma} = n^{-1}\sum_{i=1}^{n}\hat{\phi}_i\hat{\phi}_i'$ with

$$\hat{\phi}_i = W_i X_{1i}\{\bar{N}_i - m_i \exp(\hat{\beta}_1' X_{1i})\}.$$

In practice, in addition to the estimation of β, one may also be interested in checking the adequacy of models (3) and (4) given the observed data. To develop a procedure for this, define

$$\mathcal{A}(t) = \int_0^t \frac{E\{g(Z)Z\}}{E(Z)} P(C \geq u)\mu_0(u)d\Lambda_0(u),$$

and note that under models (3) and (4),

$$E\left\{\int_0^t N_i(t)d\tilde{O}_i(t)|X_i\right\} = E(m_i|X_i)\exp(\beta' X_i)\mathcal{A}(t).$$

Then, $\mathcal{A}(t)$ can be estimated by

$$\hat{\mathcal{A}}(t) = \sum_{i=1}^{n}\int_0^t \frac{N_i(u)d\tilde{O}_i(u)}{\sum_{i=1}^{n} m_i \exp(X_i'\hat{\beta})}.$$

Furthermore, for each i, define the residual

$$\hat{R}_i(t) = \int_0^t N_i(u)d\tilde{O}_i(u) - m_i \exp(X_i'\hat{\beta})\,\hat{\mathcal{A}}(t),$$

$i = 1, \ldots, n$. Then to test the goodness-of-fit of models (3) and (4), we propose to apply the statistic

$$\Phi(t, x) = n^{-1/2}\sum_{i=1}^{n} I(X_i \leq x)\,\hat{R}_i(t),$$

where the event $I(X_i \leq x)$ means that each of the components of X_i is not larger than the corresponding component of x. It is easy to see that $\Phi(t, x)$ is the cumulative sum of $\hat{R}_i(t)$ over the values of X_i's.

To apply the statistic $\Phi(t, x)$, we need to know its distribution. For this, define

$$S_0 = n^{-1}\sum_{i=1}^{n} m_i \exp(X_i'\hat{\beta}),$$

$$S(x) = n^{-1}\sum_{i=1}^{n} I(X_i \leq x) m_i \exp(X_i'\hat{\beta}),$$

and

$$B(t, x) = n^{-1}\sum_{i=1}^{n}\left\{I(X_i \leq x) - \frac{S(x)}{S_0}\right\} X_i' m_i \exp(X_i'\hat{\beta})\hat{\mathcal{A}}(t).$$

In the Appendix A.2, we will show that the null distribution of $\Phi(t, x)$ can be approximated by the zero-mean Gaussian process

$$\hat{\Phi}(t, x) = n^{-1/2}\sum_{i=1}^{n}\left\{I(X_i \leq x) - \frac{S(x)}{S_0}\right\}\hat{R}_i(t)G_i - B(t, x)' n^{-1/2}\sum_{i=1}^{n}\hat{d}_i G_i,$$

where $\hat{d}_i$ is the vector $\hat{\Gamma}^{-1}\hat{\phi}_i$ without the last entry and $(G_1, \ldots, G_n)$ are independent standard normal variables independent of the data. This suggests that one can approximate the distribution of $\Phi(t, x)$ by the empirical distribution of a large number of realizations of $\hat{\Phi}(t, x)$ given by repeatedly generating the standard normal random sample $(G_1, \ldots, G_n)$ given the observed data. Thus for checking the overall fit of models (3) and (4) based on $\Phi(t, x)$, the p-value of the omnibus test can be obtained by comparing the observed value of $\sup_{t,x}|\Phi(t, x)|$ to a large number of realizations of $\sup_{t,x}|\hat{\Phi}(t, x)|$.

Table 1
Estimation of β under the observation process (a).

$\alpha = -0.5$						
β_0	1	0	−1	1	0	−1
	$n = 100$			$n = 200$		
BIAS	0.007	−0.011	0.000	0.016	0.005	−0.002
SSE	0.299	0.296	0.311	0.213	0.210	0.211
ESE	0.267	0.265	0.277	0.205	0.206	0.201
CP	0.921	0.923	0.911	0.939	0.949	0.932
$\alpha = 0$						
BIAS	−0.011	0.005	0.005	−0.001	−0.003	0.005
SSE	0.237	0.232	0.240	0.173	0.167	0.172
ESE	0.222	0.216	0.221	0.162	0.162	0.162
CP	0.932	0.920	0.925	0.935	0.934	0.939
$\alpha = 0.5$						
BIAS	0.004	0.003	0.014	0.006	0.011	−0.004
SSE	0.198	0.186	0.215	0.137	0.145	0.135
ESE	0.176	0.176	0.199	0.131	0.144	0.131
CP	0.917	0.934	0.922	0.936	0.938	0.940

4. Simulation studies

We conducted two simulation studies to assess the performances of the proposed inference procedure with the focus on the estimation of β. The purpose of the first one was to evaluate the final sample properties of the proposed estimate, while in the second study we compared the proposed estimate to that given in Sun et al. (2007b). For the first study, the covariate X_i's were assumed to follow a Bernoulli distribution with success probability 0.5 and the latent variable Z_i's were generated from the gamma distribution with mean 10 and variance 50. Also we took $g(Z_i) = Z_i^{\alpha} + \varepsilon_i$ where $\varepsilon_i \sim$ Gamma (1, 2), and generated the follow-up time C_i from the uniform distribution over $[2, \tau + 1]$ with $\tau = 8$. It is easy to see that the parameter α represents the relationship between the observation process and the underlying recurrent event process of interest. Given the covariate X_i, $\alpha > 0$, $\alpha = 0$ and $\alpha < 0$ mean that the two processes are positively correlated, have no correlation, and are negatively correlated, respectively.

With respect to the observation process $O_i(t)$, two set-ups were considered as follows:

(a) The number of observation times m_i was assumed to follow the Poisson distribution with mean $Z_i C_i \exp(X_i)/\tau$ and the observation times $(T_{i1}, \ldots, T_{im_i})$ were taken to be the order statistics of a random sample of size m_i from the uniform distribution over $(0, C_i)$.

(b) m_i was assumed to follow the uniform distribution over $\{0, 1, 2, 3\}$ if $Z_i \leq 10$ and $\{3, 4, 5, 6\}$ otherwise, and the observation times $(T_{i1}, \ldots, T_{im_i})$ were generated in the same way as in set-up (a).

Given X_i, Z_i, m_i and $(T_{i1}, \ldots, T_{im_i})$, we generated $N_i(T_{ij})$ by using the formula

$$N_i(T_{ij}) = N_i(T_{i1}) + \{N_i(T_{i2}) - N_i(T_{i1})\} + \cdots + \{N_i(T_{ij}) - N_i(T_{i,j-1})\}$$

and assuming that $N_i(t) - N_i(s)$ followed the Poisson distribution with mean

$$\{\mu_0(t) - \mu_0(s)\}\, g(Z_i) \exp(X_i \beta_0)$$

and $\mu_0(t) = t^2/2$. All of the results given below are based on 1000 replications.

Tables 1 and 2 present the simulation results obtained on the estimation of β with the sample size $n = 100$ or 200, the true value of β equal to -1, 0 or 1, and $\alpha = 0.5$, 0, or -0.5. Table 1 corresponds to set-up (a) for the observation process and the results in Table 2 were obtained under set-up (b). Both tables include the estimated bias (BIAS) given by the average of the proposed estimates of β 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). These results indicate that the proposed estimate seems to be unbiased and the proposed variance estimation procedure provides reasonable estimates. Also the results on the empirical coverage probabilities indicate that the normal approximation seems to be appropriate.

To further investigate the robustness of the proposed estimate and also why one may need to use the proposed estimate instead of the estimates developed under restricted models such as that given in Sun et al. (2007b), we performed another simulation study to compare the estimates given here and in Sun et al. (2007b). In the study, the X_i's were assumed to follow the normal distribution with mean zero and variance 4, and we generated the latent variable Z_i's from the gamma distribution with mean 10 and variance 50. The follow-up times C_i's were generated from the uniform distribution over $[2, \tau + 1]$ with $\tau = 8$ as before. For the observation process, we assumed that $m_i \sim$ Poisson $(Z_i \Lambda_0(C_i) \exp(X_i))$ when $Z_i <= 10$ and $m_i \sim$ Poisson (4) otherwise with letting $\lambda_0(t) = 1/\tau$ or $(t + 1)/(0.5\tau^2 + \tau)$. For the case with $\lambda_0(t) = 1/\tau$, the observation times $(T_{i1}, \ldots, T_{im_i})$ were taken to be the order statistics of a random sample of size m_i from the uniform

Table 2
Estimation of β under the observation process (b).

$\alpha = -0.5$						
β_0	1	0	−1	1	0	−1
	$n = 100$			$n = 200$		
BIAS	0.002	0.024	−0.010	0.002	0.003	0.003
SSE	0.317	0.332	0.332	0.225	0.233	0.237
ESE	0.299	0.304	0.305	0.219	0.223	0.225
CP	0.931	0.931	0.929	0.946	0.933	0.938
$\alpha = 0$						
BIAS	0.002	0.006	−0.004	−0.003	0.0029	0.0061
SSE	0.295	0.302	0.310	0.196	0.216	0.207
ESE	0.273	0.272	0.273	0.196	0.197	0.199
CP	0.925	0.921	0.933	0.950	0.948	0.943
$\alpha = 0.5$						
BIAS	0.004	0.010	−0.012	0.008	−0.007	−0.006
SSE	0.236	0.252	0.257	0.177	0.179	0.173
ESE	0.236	0.235	0.241	0.172	0.172	0.173
CP	0.940	0.930	0.927	0.933	0.936	0.952

Table 3
Estimation of β based on the proposed method and STH.

$\alpha = -0.5$					
$\lambda_0(t)$	$\mu_0(t)$	$n = 100$		$n = 200$	
		Proposed	STH	Proposed	STH
$\frac{1}{\tau}$	t	−0.0045	−0.1995	0.0054	−0.1928
	$t^2/2$	−0.0359	−0.2266	−0.0122	−0.2033
$\frac{t+1}{\tau(1+0.5\tau)}$	t	−0.0073	−0.2314	−0.0060	−0.2280
	$t^2/2$	−0.0318	−0.2460	−0.0224	−0.2417
$\alpha = 0$					
$\frac{1}{\tau}$	t	0.0016	−0.2008	−0.01402	−0.2104
	$t^2/2$	−0.0358	−0.2287	−0.0235	−0.2205
$\frac{t+1}{\tau(1+0.5\tau)}$	t	−0.0096	−0.2326	−0.0116	−0.2375
	$t^2/2$	−0.0366	−0.2629	−0.0342	−0.2571
$\alpha = 0.5$					
$\frac{1}{\tau}$	t	0.0251	−0.2158	−0.0295	−0.228
	$t^2/2$	−0.0522	−0.2432	−0.0416	−0.2380
$\frac{t+1}{\tau(1+0.5\tau)}$	t	−0.0306	−0.2537	−0.0211	−0.2433
	$t^2/2$	−0.0522	−0.2783	−0.0449	−0.2773

distribution over $(0, C_i)$. For the case with $\lambda_0(t) = (t+1)/(0.5\tau^2 + \tau)$, we let the observation times $(T_{i1}, \ldots, T_{im_i})$ to be the order statistics of a random sample of size m_i from the probability density function

$$\frac{0.5t^2 + t}{0.5C_i^2 + C_i} I(0 \le t \le C_i).$$

Given the X_i's, Z_i's, m_i's and T_{ij}'s, the panel count data were generated in the same way as in the first simulation study with $\mu_0(t) = t$ or $\mu_0(t) = t^2/2$ and $g(z) = 3\log(z^\alpha/10 + 3)$. Here again α determines the correlation structure between the observation process and the underlying recurrent event process.

Table 3 gives the estimated bias (BIAS), the average of the estimates minus the true value, obtained for the two estimates of β proposed here and given in Sun et al. (2007b) based on the simulated data. The results are for $\beta_0 = 1$, $n = 100$ and 200, $\alpha = 0.5$, 0, and -0.5, and based on 1000 replications and in the table we use STH to denote the estimate proposed in Sun et al. (2007b). It is apparent that the estimate proposed here seems to be unbiased, while the estimate given in Sun et al. (2007b) is clearly biased. In other words, in general, the estimate proposed in the previous section seems to be robust.

5. An illustrative example

To illustrate the proposed inference procedure, we apply the proposed methodology to the bladder cancer data analyzed by Sun and Wei (2000) and Sun et al. among others. The data include 85 patients with bladder tumors in two treatment groups, a placebo group (47) and a thiotepa treatment group (38). During the study, each patient was observed only at discrete time points and at each time point, only the numbers of bladder tumors that had occurred since the previous observation were recorded. That is, only panel count data are available for the underlying recurrent process of bladder

Table 4
The analysis results of the bladder tumor data by different methods.

	Proposed	SW	HSW	STH	ZT
β_1	−1.3862 (0.3282)	−1.9712 (0.4423)	−1.364 (0.45)	−1.3565 (0.4387)	−1.4815 (0.3795)
β_2	0.3282 (0.0668)	0.6604 (0.2247)	0.275 (0.09)	0.2417 (0.0918)	0.2641 (0.067)
β_3	0.0000 (0.0956)	−0.1230 (0.2043)	−0.07 (0.12)	0.0134 (0.1181)	−0.0216 (0.1032)

SW: the method given in Sun and Wei (2000); HSW: Approach I with weight function 1 proposed by Hu et al. (2003); STH: the method given in Sun et al. (2007b); ZH: the method given in Zhao and Tong (2011).

tumors. For each patient, two covariates were also recorded and they are the number of initial bladder tumors and the size of the largest initial bladder tumor. Among others, Sun et al. (2007b) suggested that for the data, the underlying recurrent process and the observation process seem to be correlated. One objective of the study was to determine the treatment effect on the tumor occurrence as well as the covariate effects.

For the analysis, define $X_i = (X_{i1}, X_{i2}, X_{i3})'$ with $X_{i1} = 1$ if subject i was in the thiotepa treatment group and 0 otherwise and X_{i2} and X_{i3} denoting the number of initial tumors and the size of the largest initial tumor of the ith patient, respectively. The application of the estimation procedure proposed in the previous sections gave $\hat{\beta} = (-1.3862, 0.3282, 0.0000)'$ with the estimated standard errors of 0.3282, 0.0668 and 0.0956, 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. On the other hand, the occurrence rate did not seem to be significantly related to the size of the largest initial tumor. The data considered here were also analyzed by Sun and Wei (2000), Hu et al. (2003), Sun et al. (2007b) and Zhao and Tong (2011) among others and, for comparison, the results given by these authors are summarized in Table 4. It can be seen that the conclusion obtained here is similar to those given by others.

To check the goodness-of-fit of the models (3) and (4), we used the omnibus test procedure given in Section 3 and obtained the p-value of 0.768. This suggests that these models seem to be appropriate for the bladder cancer data considered here.

Acknowledgments

The authors are grateful to the Editor, Professor Jae Chang Lee, and an Associate Editor for their valuable comments and suggestions that greatly improved this work. The research of XZ was supported in part by the Research Grant Council of Hong Kong (PolyU 5032/09P) and The Hong Kong Polytechnic University. The research of XT was supported in part by the National Natural Science Foundation of China Zhongdian Project (11131002), the National Natural Science Foundation of China (No. 10971015), the Key Project of Chinese Ministry of Education (No. 309007) and the Fundamental Research Funds for the Central Universities.

Appendix. Proofs

In this appendix, we will sketch the proofs for the consistency and asymptotic normality of the proposed estimate $\hat{\beta}_1$ and also for the asymptotic properties of the goodness-of-fit test statistic $\Phi(t, x)$. For this, we will employ the notation defined in the previous sections and assume that $P(C \geq \tau) > 0$. Define $\Gamma = E\{WmX_1X_1' \exp(\beta_{10}'X_1)\}$ and assume that Γ is positive definite.

A.1. Proof of the asymptotic properties of $\hat{\beta}_1$

First we will consider the consistency of $\hat{\beta}_1$. For this, note the two facts:

(i) It can be easily verified that $U(\beta_{10})$ tends to 0 in probability as n tends to infinity;
(ii)

$$\frac{\partial}{\partial \beta_1} U(\beta_1) = -\frac{1}{n} \sum_{i=1}^{n} W_i m_i X_{1i} X_{1i}' \exp(X_{1i}'\beta_1)$$

converges uniformly to a negative matrix $-E\{WmX_1X_1' \exp(X_1'\beta_{10})\}$ over β_1 in a neighborhood around the true value β_{10}.

Therefore the solution $\hat{\beta}_1$ of the estimating equation $U(\beta_1) = 0$ is unique and consistent. Now we turn to prove the asymptotic normality of the proposed estimator $\hat{\beta}_1$. For this, note that by the Taylor series expansion, we have

$$n^{1/2}(\hat{\beta}_{1n} - \beta_{10}) = \Gamma^{-1} n^{-1/2} U(\beta_{10}) + o_p(1) = \Gamma^{-1} n^{-1/2} \sum_{i=1}^{n} \phi_i + o_p(1),$$

where

$$\phi_i = W_i X_{1i} \{\bar{N}_i - m_i \exp(\beta_{10}' X_{1i})\}.$$

It thus follows that $n^{1/2}(\hat{\beta}_{1n} - \beta_{10})$ has an asymptotically normal distribution with mean zero and covariance matrix $\Gamma^{-1}\Sigma(\Gamma^{-1})'$ that can be consistently estimated by $\hat{\Gamma}^{-1}\hat{\Sigma}(\hat{\Gamma}^{-1})'$, where $\Sigma = E(\phi_i\phi_i')$ and $\hat{\Gamma}$ and $\hat{\Sigma}$ are given as in Section 3.

A.2. Proof of the asymptotic property of $\Phi(t, x)$

In the following, we will sketch the proof for the weak convergence of $\Phi(t, x)$ under models (3) and (4). Assume that the limits of $S(x)$, S_0, and $B(t, x)$ exist and are denoted by $s(x)$, s_0, and $b(t, x)$, respectively. Define

$$R_i(t) = \int_0^t \left\{ N_i(u)d\tilde{O}_i(u) - m_i \exp(\beta_0' X_i)\, d\mathcal{A}(u) \right\}.$$

To prove the weak convergence of $\Phi(t, x)$, first using Lemma A.1 of Lin and Ying (2001) and the functional version of the Taylor expansion, we have

$$\Phi(t, x) = n^{-1/2} \sum_{i=1}^{n} \left\{ I(X_i \le x) - \frac{s(x)}{s_0} \right\} R_i(t) - b(t, x)' n^{1/2}(\hat{\beta} - \beta_0) + o_p(1).$$

The tightness of the first term on the right-hand side of the above follows directly from the arguments in Appendix A.5 of Lin et al. (2000). The second term is also tight because $n^{1/2}(\hat{\beta} - \beta_0)$ converge in distribution and $b(t, x)$ is a deterministic function. Thus $\Phi(t, x)$ is tight.

Let d_i be the vector $\Gamma^{-1}\phi_i$ without the last entry. Then, we can further write $\Phi(t, x)$ as

$$\Phi(t, x) = n^{-1/2} \sum_{i=1}^{n} \left\{ I(X_i \le x) - \frac{s(x)}{s_0} \right\} R_i(t) - b(t, x)' n^{-1/2} \sum_{i=1}^{n} d_i + o_p(1).$$

It thus follows from the multivariate central limit theorem and the tightness of $\Phi(t, x)$ that $\Phi(t, x)$ converges weakly to a zero-mean Gaussian process that can be approximated by the zero-mean Gaussian process

$$\tilde{\Phi}(t, x) = n^{-1/2} \sum_{i=1}^{n} \left\{ I(X_i \le x) - \frac{S(x)}{S_0} \right\} \hat{R}_i(t) - B(t, x)' n^{-1/2} \sum_{i=1}^{n} \hat{d}_i.$$

Thus, using the simulation approach presented in Lin et al. (2000), the null distribution of $\Phi(t, x)$ can be approximated by that of $\hat{\Phi}(t, x)$.

References

Balakrishnan, N., Zhao, X., 2009. New multi-sample nonparametric tests for panel count data. Annals of Statistics 37, 1112–1149.
Cai, J., Schaubel, D.E., 2004. Analysis of recurrent event data. Handbook of Statistics 23, 603–623.
Cheng, S.C., Wei, L.J., 2000. Inferences for a semiparametric model with panel data. Biometrika 87, 89–97.
Cook, R.J., Lawless, J.F., 2007. The Statistical Analysis of Recurrent Events. Springer-Verlag, New York.
Diggle, P.J., Liang, K.Y., Zeger, S.L., 1994. The Analysis of Longitudinal Data. Oxford University Press, Oxford.
Gibbons, J.D., Chakraborti, S., 2011. Nonparametric Statistical Inference, fifth ed. Chapman & Hall.
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.
Hu, X.J., Lagakos, S.W., Lockhart, R.A., 2009. Generalized least squares estimation of the mean function of a counting process based on panel counts. Statistica Sinica 19, 561–580.
Hu, X.J., Sun, J., Wei, L.J., 2003. Regression parameter estimation from panel counts. Scandinavian Journal of Statistics 30, 25–43.
Liang, Y., Lu, W., Ying, Z., 2009. Joint modeling and analysis of longitudinal data with informative observation times. Biometrics 65, 377–384.
Lin, D.Y., Wei, L.J., Yang, I., Ying, Z., 2000. Semiparametric regression for the mean and rate function of recurrent events. Journal of the Royal Statistical Society B 69, 711–730.
Lin, D.Y., Ying, Z., 2001. Semiparametric and nonparametric regression analysis of longitudinal data. Journal of the American Statistical Association 96, 103–126.
Liu, L., Huang, X., O'Quigley, J., 2008. Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. Biometrics 64, 950–958.
Lu, M., Zhang, Y., Huang, J., 2007. Estimation of the mean function with panel count data using monotone polynomial splines. Biometrika 94, 1–14.
Lu, M., Zhang, Y., Huang, J., 2009. Semiparametric estimation methods for panel count data using monotone *B*-splines. Journal of the American Statistical Association 104, 1060–1070.
Sun, J., 2006. The Statistical Analysis of Interval–Censored Failure Time Data. Springer, 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., Sun, L., Liu, D., 2007a. Regression analysis of longitudinal data in the presence of informative observation and censoring times. Journal of the American Statistical Association 102, 1397–1406.
Sun, J., Tong, X., He, X., 2007b. 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 B 62, 293–302.
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.
Wellner, J.A., Zhang, Y., 2007. Two likelihood-based semiparametric estimation methods for panel count data with covariates. Annals of Statistics 35, 2106–2142.
Welsh, A.H., Lin, X., Carroll, R.J., 2002. Marginal longitudinal nonparametric regression: locality and efficiency of spline and kernel methods. Journal of the American Statistical Association 97, 482–493.
Zhang, Y., 2006. Nonparametric *k*-sample tests for panel count data. Biometrika 93, 777–790.
Zhang, Y., 2002. A semiparametric pseudolikelihood estimation method for panel count data. Biometrika 89, 39–48.
Zhang, Y., Jamshidian, M., 2003. The gamma-frailty Poisson model for the nonparametric estimation of panel count data. Biometrics 59, 1099–1106.
Zhao, X., Tong, X., 2011. Semiparametric regression analysis of panel count data with information observation times. Computational Statistics and Data Analysis 55, 291–300.