Computational Statistics and Data Analysis 55 (2011) 291–300

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Semiparametric regression analysis of panel count data with informative observation times

Xingqiu Zhao [a], Xingwei Tong [b,*]

[a] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong

[b] School of Mathematical Sciences, Beijing Normal University, Beijing, PR China

ARTICLE INFO

Article history:
Received 3 October 2009
Received in revised form 5 March 2010
Accepted 25 April 2010
Available online 16 May 2010

Keywords:
Estimating equation
Informative observation times
Mean function model
Panel count data
Regression analysis

ABSTRACT

This paper discusses regression analysis of panel count data that arise naturally when recurrent events are considered. For the analysis of panel count data, most of the existing methods have assumed that observation times are completely independent of recurrent events or given covariates, which may not be true in practice. We propose a joint modeling approach that uses an unobserved random variable and a completely unspecified link function to characterize the correlations between the response variable and the observation times. For inference about regression parameters, estimating equation approaches are developed without involving any estimation for latent variables, and the asymptotic properties of the resulting estimators are established. In addition, a technique is provided for assessing the adequacy of the model. The performance of the proposed estimation procedures are evaluated by means of Monte Carlo simulations, and a data set from a bladder tumor study is analyzed as an illustrative example.

1. Introduction

The analysis of panel count data has recently attracted considerable attention. Panel count data include the number of observations, discrete observation times, the counts of recurrent events and the censoring or follow-up times for each study subject. Furthermore, both observation and follow-up times may vary from subject to subject. Such data frequently occur in medical periodic follow-up studies, reliability experiments, AIDS clinical trials, animal tumorgenicity experiments, and sociological studies (Kalbfleisch and Lawless, 1985; Thall and Lachin, 1988).

For the analysis of panel count data, several nonparametric and semiparametric methods have been developed, and 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 assuming that the counting process is independent of the number of observations and the observation times, while Sun and Fang (2003), Zhang (2006), and Balakrishnan and Zhao (2009, 2010) presented nonparametric tests for the problem of nonparametric comparison of the mean function of counting processes with panel count data. Zhang and Jamshidian (2003) introduced the gamma frailty variable to account for correlation among the panel counts and used the maximum pseudo-likelihood approach to estimate the mean function. Sun and Wei (2000) proposed a semiparametric regression model for the mean function of the cumulative number of recurrent events over time and used estimating equation-based methods when both observation times and the censoring times may depend on covariates. Zhang (2002) and Wellner and Zhang (2007) discussed

* Corresponding author.
E-mail address: xweitong@bnu.edu.cn (X. Tong).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.04.020

regression analysis of panel count data by using the semiparametric likelihood-based approach under the assumption that the observation times are independent of occurrences of the recurrent event under study given covariates. Lu et al. (2007) studied nonparametric likelihood-based estimators of the mean function of counting processes with panel count data, using monotone polynomial splines. Cheng and Wei (2000) investigated estimating equation approaches when observation and censoring times are independent of the event process. Hu et al. (2003) also investigated estimating equation approaches for the case when the event process, the observation process, and the censoring time are independent conditional on covariates.

In practice, 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 by Sun and Wei (2000), Huang et al. (2006) and Sun et al. (2007). In the AIDS study described in Wang et al. (2001), suppose that one is interested in some symptoms related to AIDS, such as CD4 counts or the time at which the patient's CD4 counts cross some threshold. Then the response process may be correlated with the observation process. Lipsitz et al. (2002) presented a set of longitudinal data from a study of children with acute lymphoblastic leukemia which involves correlated response and observation processes. The same could be appropriate for other medical follow-up studies. Thus, one could have to deal with two related processes. 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. The random effects are used to link the two processes. Huang et al. (2006) studied nonparametric and semiparametric models that allow observation times to be correlated with the event process through a frailty variable, and used the conditional likelihood approach to estimate the baseline function and the regression parameters. Furthermore, Sun et al. (2007) investigated semiparametric models for the observation process and the 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. However, we notice that the estimators obtained from the estimating equation approach proposed by Sun et al. (2007) are not consistent and they do not have asymptotic normality. The reason for this will be discussed in the next section. The main objective of this paper is to consider more general joint models for panel count data with informative observation times and develop an estimating equation approach for estimation of regression parameters such that the asymptotic properties of the resulting estimates can be established.

The remainder of this paper is organized as follows. Section 2 introduces notation and describes joint models for the recurrent event and observation processes, where an unobserved random variable and a completely unspecified link function are used to characterize the correlation between the two processes. In Section 3, estimating equation approaches are proposed for estimation of regression parameters and the asymptotic properties of the resulting estimates are established. In Section 4, we develop a technique for checking the adequacy of the general joint model. Section 5 presents some results obtained from simulation studies for assessing the finite-sample properties of the proposed inference procedure and comparing it with some existing methods in the literature. In Section 6, we apply the proposed methods to a data set from a bladder tumor study. Some concluding remarks are made in Section 7.

2. Joint 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

$$E\{N_i(t)|x_i, z_i\} = \mu_N(t)g(z_i)\exp(x_i'\beta). \tag{1}$$

Here $\mu_N(\cdot)$ is a completely unknown continuous baseline mean function, $g(\cdot)$ is a completely unspecified function, and β is a vector of unknown regression parameters.

For subject i, suppose that $N_i(\cdot)$ is observed only at finite time points $T_{ij} < \cdots < T_{iK_i}$, 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 be the censoring time and thus $N_i(t)$ is observed only at these T_{ij}'s with $T_{ij} \le C_i$, $i = 1, \ldots, n$. Define $\tilde{H}_i(t) = H_i\{\min(t, C_i)\}$, where $H_i(t) = \sum_{j=1}^{K_i} I(T_{ij} \le t)$, $i = 1, \ldots, n$. Then $\tilde{H}_i(t)$ is a point process characterizing the ith subject's observation process and jumps only at the observation times.

In the following, we assume that given x_i and z_i, $H_i(\cdot)$ is a non-homogeneous Poisson process with the intensity function

$$\lambda_h(t|x_i, z_i) = \lambda_{0h}(t)z_i\exp(x_i'\gamma). \tag{2}$$

In model (2), $\lambda_{0h}(t)$ is a completely unknown continuous baseline intensity function and γ denotes the vector of regression parameters. Let $\Lambda_{0h}(t) = \int_0^t \lambda_{0h}(s)ds$. We assume that $\Lambda_{0h}(\tau) = 1$ for identifiability, where τ denotes the length of study. In addition, we assume that N_i's and H_i's are independent conditional on (x_i, z_i), C_i's are independent of (N_i, H_i, x_i, z_i)'s, and $\{H_i(t), N_i(t), C_i, x_i', u_i, z_i, 0 \le t \le \tau\}$, $i = 1, \ldots, n$, are independent and identically distributed.

There exists a great deal of research on each of the two models (1) and (2) and their special cases individually. For example, model (1) with $g(z_i) = 1$ was considered by Sun and Wei (2000), Zhang (2002) and Wellner and Zhang (2007) for regression analysis of panel count data; model (2) with $z_i = 1$ is commonly used for regression analysis of recurrent event data.

Wang et al. (2001) and Huang and Wang (2004) considered a model similar to model (2) for recurrent event data. In contrast, there exists limited work on the joint analysis of the two models. For the joint models, Sun et al. (2007) studied a special case by assuming that $g(z_i) = z_i^{\alpha}$ where α is an unknown parameter, and provided an inference procedure by using a moment-based estimator $\hat{z}_i$ instead of z_i in the estimating function for β. However, the estimators for regression parameters obtained in such way are not consistent because the estimating function involves estimation of $z_i^{1+\alpha}$, which is nonlinear on z_i. In the following, we study the joint analysis of the two models together with the focus on estimation of regression parameters β along with γ. Conditional on x_i, the correlation of two processes $N(\cdot)$ and $H(\cdot)$ is characterized by a completely unknown function of z_i, while the distributional form of z_i is left unspecified.

3. Estimation of regression parameters

In this section, we consider estimation of β along with other parameters. For this, note that if the random effects z_i's are known, then model (1) becomes the usual proportional means model and several methods, such as that given in Cheng and Wei (2000), can be used. Unfortunately they are unknown in practice. To deal with this, following Sun and Wei (2000) and Sun et al. (2007), we define

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

where $K_i^* = \tilde{H}_i(C_i)$, the total number of observations on subject i, $i = 1, \ldots, n$. Then conditional on x_i, we have

$$E\left(\bar{N}_i|x_i\right) = \exp\{x_i'(\beta+\gamma)\} E\{g(z_i)z_i\} \int_0^{\tau} \lambda_{0h}(t) P(C_i \ge t) \mu_N(t) \mathrm{d}t. \tag{3}$$

For estimation of β, motivated by Eq. (3), one can use the following estimating function

$$U(\beta_1|\gamma) = \frac{1}{n} \sum_{i=1}^{n} x_{1i} \left\{\bar{N}_i - \exp(x_{1i}'\beta_1 + x_i'\gamma)\right\},$$

where $x_{1i}' = (x_i', 1)$ and $\beta_1' = (\beta', \nu)$ with

$$\nu = \log\left[E\{g(z_i)z_i\} \int_0^{\tau} \lambda_{0h}(t) P(C_i \ge t) \mu_N(t) \mathrm{d}t\right].$$

The $U(\beta_1|\gamma)$ is an unbiased estimating function, that is, the expected value of $U(\beta_1|\gamma)$ is 0. Thus, it is natural to estimate β by the solution to $U(\beta_1|\gamma) = 0$.

Of course, γ is unknown and $U(\beta_1|\gamma)$ is not available. For this, we first consider inference about model (2), for which we have recurrent event data. Let the s_l's denote the ordered and distinct time points of all the observation times $\{T_{ij}\}$, d_l the number of the observation times equal to s_l, and n_l the number of the observation times satisfying $T_{ij} \le s_l \le C_i$ among all subjects. Set $\gamma_*' = (\gamma', \gamma_1)$, where $\gamma_1 = \log\{E(z_i)\}$. Then following Huang and Wang (2004), one can first estimate $\Lambda_{0h}(t)$ and γ_* by

$$\widehat{\Lambda}_{0h}(t) = \prod_{s_l > t} \left(1 - \frac{d_l}{n_l}\right)$$

and the estimating equation

$$\sum_{i=1}^{n} x_{1i} \left\{K_i^* \widehat{\Lambda}_{0h}^{-1}(C_i) - \exp(\gamma_*' x_{1i})\right\} = 0, \tag{4}$$

respectively. Denote the solution to Eq. (4) as $\hat{\gamma}_n^*$. A key fact used in deriving the above estimating equation is that conditional on (x_i, C_i, z_i, K_i^*), the observation times $\{T_{i1}, \ldots, T_{iK_i^*}\}$ are the order statistics of a simple random sample of size K_i^* from the density function

$$\frac{\lambda_{0h}(t) z_i \exp(\gamma' x_i)}{\Lambda_{0h}(C_i) z_i \exp(\gamma' x_i)} I(0 \le t \le C_i) = \frac{\lambda_{0h}(t)}{\Lambda_{0h}(C_i)} I(0 \le t \le C_i).$$

Let $\hat{\gamma}_n$ be $\hat{\gamma}_n^*$ without the last entry. Then we can use the estimating function $U(\beta_1|\hat{\gamma}_n)$ to estimate regression parameters in model (1). Let $\hat{\beta}_{1n}$ denote the solution to equation $U(\beta_1|\hat{\gamma}_n) = 0$ and let $\beta_{10} = (\beta_0', \nu)'$ be the true value of β_1. To establish the asymptotic properties of $\hat{\beta}_{1n}$, we need the following regularity conditions, which are similar to those given in Huang and Wang (2004).

C.1. $P(C \ge \tau, Z > 0) > 0$.
C.2. X is uniformly bounded.
C.3. $G(s) = E\{ZI(C \ge s)\}$ is continuous for $s \in [0, \tau]$.

Theorem 1. *Assume that Conditions* C.1–C.3 *hold. Then, the estimator* $\hat{\beta}_{1n}$ *of* β_{10} *is consistent and* $\sqrt{n}(\hat{\beta}_{1n} - \beta_{10})$ *has an asymptotic normal distribution with mean zero and the covariance matrix* $\phi^{-1}\Sigma(\phi^{-1})'$, *where* ϕ *and* Σ *are given in the Appendix.*

The proof of the above theorem is sketched in the Appendix. To derive the estimation of the covariance, we introduce some notation. Let

$$Q_n(t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K_i^*} I(T_{ij} \le t),$$

$$R_n(t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K_i^*} I(T_{ij} \le t \le C_i),$$

and

$$b_{in}(t) = \sum_{j=1}^{K_i^*}\left\{\int_t^{\tau}\frac{I(T_{ij} \le u \le C_i)}{R_n^2(u)}\mathrm{d}Q_n(u) - \frac{I(t \le T_{ij} \le \tau)}{R_n(T_{ij})}\right\}$$

for $t \in [0, \tau]$ and $i = 1, \ldots, n$. Also define

$$\hat{f}_{ih} = -\frac{1}{n}\sum_{j=1}^{n} x_{1j}K_j^* b_{in}(C_j)\{\hat{\Lambda}_{0h}(C_j)\}^{-1} + x_{1i}[K_i^*\{\hat{\Lambda}_{0h}(C_i)\}^{-1} - \exp(x_{1i}'\hat{\gamma}_n^*)]$$

and define $\hat{f}_i$ as the vector $\{\frac{1}{n}\sum_{j=1}^{n} x_{1j}x_{1j}'\exp(x_{1j}'\hat{\gamma}_n^*)\}^{-1}\hat{f}_{ih}$ without the last entry.

Then ϕ and Σ can be, respectively, consistently estimated by

$$\hat{\phi} = -\frac{1}{n}\sum_{i=1}^{n} x_{1i}x_{1i}'\exp(x_{1i}'\hat{\beta}_{1n} + x_i'\hat{\gamma}_n),$$

and

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\left[x_{1i}\left\{\bar{N}_i - \exp(x_{1i}'\hat{\beta}_{1n} + x_i'\hat{\gamma}_n)\right\} - \frac{1}{n}\sum_{j=1}^{n}\left\{x_{1j}\exp(x_{1j}'\hat{\beta}_{1n} + x_j'\hat{\gamma}_n)x_j'\right\}\hat{f}_i\right]^{\otimes 2},$$

where $a^{\otimes 2} = aa'$ for any vector a.

4. Model diagnostics

For the checking of model (2), one has complete recurrent event data and can find some approaches presented in Huang and Wang (2004). Here we consider the assessment of model (1) with model (2), and describe some graphical and numerical procedures for checking the adequacy of the proposed joint model. Let

$$\mathcal{A}(t) = \int_0^t E\{g(Z)Z\}\mu_N(u)\mathrm{d}\Lambda_0(u),$$

which can be estimated by

$$\hat{\mathcal{A}}(t) = \sum_{i=1}^{n}\int_0^t \frac{\Delta_i(u)N_i(u)\mathrm{d}H_i(u)}{\sum_{i=1}^{n}\Delta_i(u)\exp(x_i'\hat{\theta})},$$

where $\Delta_i(u) = I(C_i \ge u)$ and $\hat{\theta} = \hat{\beta}_n + \hat{\gamma}_n$, with $\hat{\beta}_n$ being $\hat{\beta}_{1n}$ without the last entry. For each i, following Lin et al. (2000), we define the residual

$$\hat{M}_i(t) = \int_0^t\left[\Delta_i(u)N_i(u)\mathrm{d}H_i(u) - \Delta_i(u)\exp(x_i'\hat{\theta})\mathrm{d}\hat{\mathcal{A}}(u)\right],$$

$i = 1, \ldots, n$. First, we consider checking the functional form for the kth component of X and we may plot $\hat{M}_i(t)$ against x_{ik}, where x_{ik} is the kth component of x_i. To develop a more formal procedure, we let

$$\Phi_k(x) = n^{-1/2}\sum_{i=1}^{n} I(x_{ik} \le x)\hat{M}_i(\tau),$$

which is the cumulative sum of $\hat{M}_i(\tau)$ over the values of x_{ik}. Define

$$S_0(t) = n^{-1}\sum_{i=1}^{n} \Delta_i(t)\exp(x_i'\hat{\theta}),$$

$$S_k(t,x) = n^{-1}\sum_{i=1}^{n} I(x_{ik}\le x)\Delta_i(t)\exp(x_i'\hat{\theta}),$$

and

$$B(t,x) = n^{-1}\sum_{i=1}^{n}\int_0^t \left\{I(x_{ik}\le x) - \frac{S_k(u,x)}{S_0(u)}\right\} x_i\Delta_i(u)\exp(x_i'\hat{\theta})\mathrm{d}\hat{\mathcal{A}}(u).$$

To use the statistic $\Phi_k(x)$, we show in the Appendix that its null distribution can be approximated by the zero-mean Gaussian process

$$\tilde{\Phi}_k(x) = n^{-1/2}\sum_{i=1}^{n}\int_0^\tau \left\{I(x_{ik}\le x) - \frac{S_k(u,x)}{S_0(u)}\right\}\mathrm{d}\hat{M}_i(u) - B(\tau,x)'n^{-1/2}\sum_{i=1}^{n}(\hat{f}_i+\hat{g}_i), \quad (5)$$

where $\hat{g}_i$ is the vector $\hat{\phi}^{-1}U_i(\hat{\beta}_{1n})$ without the last entry.

Following the simulation approach presented in Lin et al. (2000), we let $(G_1,\ldots,G_n)$ be independent standard normal variables independent of the data. Then it can be shown that the distribution of the process $\Phi_k(x)$ can be approximated by that of the zero-mean Gaussian process

$$\hat{\Phi}_k(x) = n^{-1/2}\sum_{i=1}^{n}\int_0^\tau \left\{I(x_{ik}\le x) - \frac{S_k(u,x)}{S_0(u)}\right\}\mathrm{d}\hat{M}_i(u)G_i - B(\tau,x)'n^{-1/2}\sum_{i=1}^{n}(\hat{f}_i+\hat{g}_i)G_i. \quad (6)$$

It follows from (5) and (6) that to approximate the distribution of $\Phi_k(x)$, one can obtain a large number of realizations from $\hat{\Phi}_k(x)$ by repeatedly generating the standard normal random sample $(G_1,\ldots,G_n)$ given the observed data. To assess the functional form of the jth component of covariates, one can plot a few realizations from $\hat{\Phi}_k(x)$ along with the observed $\Phi_k(x)$ and see if they can be regarded as arising from the same population. More formally, we can apply the supremum test statistic $\sup_x|\Phi_k(x)|$, where the p-value can be obtained by comparing the observed value of $\sup_x|\Phi_k(x)|$ to a large number of realizations from $\sup_x|\hat{\Phi}_k(x)|$.

An omnibus test for checking the overall fit of model (1) with model (2) can be constructed from the process

$$\Phi_0(t,x) = n^{-1/2}\sum_{i=1}^{n} I(x_i\le x)\hat{M}_i(t),$$

where the event $I(x_i\le x)$ means that each of the components of x_i is no larger than the corresponding component of x. As with $\Phi_k(x)$, we can similarly show that the null distribution of $\Phi_0(t,x)$ can be approximated by that of the zero-mean Gaussian process $\hat{\Phi}_0(t,x)$, which is obtained from the expression (6) by replacing $I(x_{ik}\le x)$ with $I(x_i\le x)$, τ in the first integral with t, and $B(\tau,x)$ with $B(t,x)$. An omnibus test statistic is then given by $\sup_{t,x}|\Phi_0(t,x)|$, where the p-value can be obtained as that of $\sup_x|\Phi_k(x)|$.

5. Simulation study

First we conducted a simulation study to assess the performance of the estimation procedure proposed in the previous sections under different situations. In the study, the covariate x_i's were assumed to follow a Bernoulli distribution with success probability 0.5. To generate the simulated data, we first generated z_i from the gamma distribution with mean 10 and variance 50, $g(z_i) = z_i^\alpha + \Gamma(1,2)$, and the follow-up time C_i from the uniform distribution on $[\tau/2,\tau]$ with $\tau = 18$, respectively. Here the symbol of α characterizes the relationship between the observation process and the recurrent event process. 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 H_i follows the homogeneous Poisson process with $\lambda_{0h}(t) = \tau^{-1}$. Then given x_i, z_i, and C_i, K_i^*, the number of real observation times for subject i, follows the Poisson distribution with mean

$$\Lambda_h(C_i|x_i,z_i) = \Lambda_{0h}(C_i)z_i\exp(x_i\gamma) = \frac{C_iz_i\exp(x_i\gamma)}{\tau},$$

$i = 1,2,\ldots,n$. Furthermore, the observation times $(T_{i1},\ldots,T_{iK_i^*})$ are the order statistics of a random sample of size K_i^* from the uniform distribution over $(0,C_i)$. Given K_i^* and $(T_{i1},\ldots,T_{iK_i^*})$, we generated $N_i(T_{ij})$ using the formula

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

Table 1
Estimation of β under the homogeneous Poisson observation process.

	$\alpha = -0.5$: H and N are negatively correlated					
β_0	1	0	−1	1	0	−1
	$n = 100$			$n = 200$		
$\hat{\beta}$	0.9965	−0.0107	−0.9912	1.0060	0.0067	−1.0030
SSD	0.2716	0.2874	0.2832	0.1663	0.1709	0.2107
ESD	0.2654	0.2652	0.2690	0.1663	0.1687	0.2012
CP	0.9360	0.9240	0.9310	0.9460	0.9410	0.9430
	$\alpha = 0$: H and N have no correlation					
β_0	1	0	−1	1	0	−1
	$n = 100$			$n = 200$		
$\hat{\beta}$	0.9961	−0.0004	−0.9994	0.9985	0.0024	−0.9935
SSD	0.2443	0.2392	0.2444	0.1698	0.1749	0.1723
ESD	0.2244	0.2234	0.2294	0.1644	0.1658	0.1679
CP	0.9290	0.9210	0.9240	0.9410	0.9320	0.9320
	$\alpha = 0.5$: H and N are positively correlated					
β_0	1	0	−1	1	0	−1
	$n = 100$			$n = 200$		
$\hat{\beta}$	1.0080	0.0021	−0.9899	1.0030	0.0094	−1.0080
SSD	0.1982	0.1978	0.1979	0.1389	0.1396	0.1372
ESD	0.1851	0.1878	0.1882	0.1365	0.1365	0.1380
CP	0.9280	0.9300	0.9370	0.9340	0.9380	0.9550

for $j = 1, \ldots, K_i^*$ and $i = 1, \ldots, n$. Here the random variable $N_i(t) - N_i(s)(t > s)$ can be generated from the Poisson distribution with mean

$$0.5(t^2 - s^2)g(z_i)\exp(x_i\beta).$$

Set $\gamma = 1$ and $\beta = -1, 0, 1$, representing the different effect of the covariate X on the panel count. For each setting, we consider two sample sizes, $n = 100$ and 200, respectively. We also performed Monte Carlo studies when the observation process H_i follows the nonhomogeneous Poisson process with $\lambda_{0h}(t) = (t+1)/(\tau(\tau/2+1))$. All the results reported here are based on 1000 Monte Carlo replications using MATLAB software.

Tables 1 and 2 present the simulation results on estimation of β for the different situations. The tables include the averages of proposed estimates of β based on the simulated data, the sample standard deviations of the estimates (SSD), the means of the estimated standard deviations (ESD), and the empirical 95% coverage probabilities (CP) for β. These results indicate that the estimate $\hat{\beta}$ 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.

Also we carried out a simulation study to compare the performance of the proposed estimator of β with those obtained by Hu et al. (2003) and Sun et al. (2007). Let $\hat{\beta}_{\mathrm{HSW}}$ denote the estimator of β based on the estimation conditional on the observational process, and let $\hat{\beta}_{\mathrm{STH}}$ denote the estimator of β presented by Sun et al. (2007). We use the efficiency to measure the gain by our joint model, where the efficiency is defined by $\mathrm{Eff} = \mathrm{SSD}(\hat{\beta}_{\mathrm{HSW}})/\mathrm{SSD}(\hat{\beta})$. For simplicity, we assume that $\mu_N(t) = t$, z_i's follow the uniform distribution $U[0, 1]$, and $g(z) = z^{\alpha}$. Since the two estimators are both unbiased, we just presented the efficiency results. The simulation results are shown in Tables 3 and 4. The two tables show that all the efficiencies are greater than one, which indicates that our estimators are more efficient than those obtained by Hu et al. (2003) and Sun et al. (2007).

6. An application

In this section, we use the proposed methodology to analyze the data set from a bladder cancer study conducted by the Veterans Administration Cooperative Urological Research Group (Andrews and Herzberg, 1985; Byar, 1980; Sun and Wei, 2000; Wellner and Zhang, 1998). In this study, the patients with superficial bladder tumors were randomly assigned to one of three treatment groups: placebo, thiotepa, and pyridoxine. Following Sun and Wei (2000), we restrict our attention to the patients in the placebo (47) and the thiotepa (38) groups. For each patient, two other important baseline covariates were reported; they are the number of initial tumors and the size of the largest initial tumor.

For the analysis, we define the first component x_{i1} of x_i to be equal to 1 if the ith patient was given the thiotepa treatment and 0 otherwise. We also let x_{i2} and x_{i3} be the number of initial tumors and the size of the largest initial tumor of the ith patient, respectively. Assume that the occurrence process of the bladder tumors and the clinical visit process can be described by models (1) and (2), respectively. The application of the estimation procedure proposed in the previous sections

Table 2
Estimation of β under the nonhomogeneous Poisson observation process.

	$\alpha = -0.5$: H and N are negatively correlated					
β_0	1	0	−1	1	0	−1
	$n = 100$			$n = 200$		
$\hat{\beta}$	1.0200	0.0166	−1.0010	1.0160	0.0159	−0.9997
SSD	0.2715	0.2780	0.2608	0.1874	0.1963	0.1956
ESD	0.2494	0.2517	0.2515	0.1831	0.1837	0.1843
CP	0.9210	0.9270	0.9420	0.9340	0.9380	0.9520
	$\alpha = 0$: H and N have no correlation					
β_0	1	0	−1	1	0	−1
	$n = 100$			$n = 200$		
$\hat{\beta}$	1.0170	0.0184	−0.9825	1.0240	0.0112	−0.9863
SSD	0.2524	0.2606	0.2608	0.1835	0.1844	0.1926
ESD	0.2422	0.2422	0.2430	0.1772	0.1773	0.1773
CP	0.9290	0.9220	0.9260	0.9400	0.9300	0.9350
	$\alpha = 0.5$: H and N are positively correlated					
β_0	1	0	−1	1	0	−1
	$n = 100$			$n = 200$		
$\hat{\beta}$	1.0160	0.0033	−0.9779	1.0040	0.0053	−0.9911
SSD	0.2387	0.2438	0.2446	0.1705	0.1653	0.1737
ESD	0.2223	0.2240	0.2229	0.1651	0.1647	0.1653
CP	0.9260	0.9250	0.9340	0.9400	0.9460	0.9310

Table 3
Efficiency results of the proposed estimators with those by Hu et al. (2003).

β_0	1	0	−1	1	0	−1
α	$n = 100$			$n = 200$		
−0.5	1.1361	1.0482	1.1848	1.1946	1.1765	1.2550
0	1.2067	1.0664	1.0228	1.0941	1.0111	1.0156
0.5	1.0582	1.0912	1.1866	1.2636	1.0908	1.1058

Table 4
Efficiency results of the proposed estimators with those by Sun et al. (2007).

β_0	1	0	−1	1	0	−1
α	$n = 100$			$n = 200$		
−0.5	1.4444	1.4297	1.381	1.4509	1.5452	1.4122
0	1.1624	1.1748	1.1514	1.2409	1.2285	1.1771
0.5	1.0557	1.1046	1.1063	1.0993	1.0694	1.1033

Table 5
Estimates and their estimated standard error by different methods.

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

SW: proposed by Sun and Wei (2000); HSW: Approach I with weight function 1 proposed by Hu et al. (2003); STH: proposed by Sun et al. (2007).

gave $\hat{\gamma}_1 = 0.4808$, $\hat{\gamma}_2 = -0.0358$, $\hat{\gamma}_3 = 0.0155$, $\hat{\beta}_1 = -1.4815$, $\hat{\beta}_2 = 0.2641$, and $\hat{\beta}_3 = -0.0216$ with the estimated standard errors being 0.1231, 0.0352, 0.0377, 0.3795, 0.0670 and 0.1032, which correspond to p-values of 0.0001, 0.3091, 0.6810, 0.0001, 0.0001, and 0.8345, respectively. Here γ_1 and β_1, γ_2 and β_2, and γ_3 and β_3 represent regression coefficients corresponding to the treatment indicator, the number of initial tumors, and the size of the largest initial tumor, respectively. These results suggest that the thiotepa treatment significantly reduces the occurrence rate of the bladder tumors and the number of initial tumors has a significant positive effect on the tumor recurrence rate but no significant effect on the visit process. However, both the occurrence rate of the bladder tumors and the visit times do not seem to be significantly related to the size of the largest initial tumor. Sun and Wei (2000) and Hu et al. (2003) also analyzed the same data. Their results are summarized in Table 5. In addition, the estimation results are shown in Table 5 by using the approach of Sun et al. (2007) to the these data. From Table 5, one can obtain similar conclusions from all the four methods. Furthermore, one can see that our proposed approach yields the smallest sample standard deviations and this means that our approach has the best efficiency.

We also use the proposed model checking techniques described in the previous section to check the functional form for the kth ($k = 1, 2, 3$) component of X and the overall fit of model (1) with model (2). Using the approaches, we computed the test statistics and their corresponding p-values. The p-values are: 0.930 for $\sup_x |\Phi_1(x)|$, 0.855 for $\sup_x |\Phi_2(x)|$, 0.786 for $\sup_x |\Phi_3(x)|$, and 0.952 for $\sup_{t,x} |\Phi_0(t, x)|$. All the p-values are significantly greater than 0.05, which shows that our proposed models are appropriate to analyze the bladder cancer data.

7. Concluding remarks

In this article, we have considered regression analysis of panel count data when the observation times may carry information about the recurrent event process and proposed more general joint models, where the relation between the two processes is characterized through a latent variable and a completely unspecified link function. For estimation of regression parameters representing covariate effects, we have developed an estimating equation approach that yields consistent and asymptotically normal parameter estimates. A key advantage of the proposed approach over existing methods for panel count data is that it allows the observation process to be related with the response process of interest through an unspecified link function of an unobserved random variable, while the link function in the joint model considered by Sun et al. (2007) is specified. Another advantage of the proposed method is that the parameter estimates and the estimated covariance matrix do not involve estimation of the latent variables, while the estimating function proposed by Sun et al. (2007) does involve estimation of the latent variables and their estimation procedure does not lead to consistency of estimators for regression parameters. The simulation results suggest that the proposed inference procedures perform well and an illustrative example is provided.

Note that in the foregoing we have assumed that the follow-up or censoring time is independent of covariates for simplicity of presentation, and the proposed method can be generalized to the dependent case by using the approach of Sun et al. (2007). Also note that we can use the weighted estimating function for inference, and it would be of great interest to investigate the weight function selection for optimal estimates.

Although the methods proposed by Sun and Wei (2000) and Hu et al. (2003) are for the case of noninformative observation times, the estimating function for β given in Section 4 of Sun and Wei (2000) and Approach II given in Section 2.3 of Hu et al. (2003) can be extended to the case of informative observation times. That is, to estimate β, motivated by (3) and Sun and Wei (2000) as well as Hu et al. (2003), we propose two other estimating functions

$$U_1(\beta|\gamma) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \exp\{-x_i'(\beta + \gamma)\}\bar{N}_i$$

or

$$U_2(\beta|\gamma) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} I(C_i \geq t)\left[x_i - \frac{\sum_{k=1}^{n} I(C_k \geq t)x_k \exp\{x_k'(\beta + \gamma)\}}{\sum_{k=1}^{n} I(C_k \geq t) \exp\{x_k'(\beta + \gamma)\}}\right] N_i(t)\mathrm{d}H_i(t),$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Clearly, the estimators of β obtained from $U_1(\beta|\hat{\gamma}) = 0$ and $U_2(\beta|\hat{\gamma}) = 0$ have the same asymptotic properties as those given in Sun and Wei (2000) and Hu et al. (2003), respectively. A simulation study indicates that estimators of β have similar performance by using the three estimating equations. In summary, we have three methods available for estimation of β under the joint model (1) and (2) and all of them work well.

It is of main interest in this article to estimate the regression parameter β from the panel count data with the observed covariates. It is hard to estimate the baseline mean function $\mu_N(t)$ in the current setting. Further research is needed to address this issue.

In the joint models, we have assumed that the covariates and the latent variables are time-independent. In some situations, one may want to consider the joint models for the recurrent event process $N(t)$ and the observation process $H(t)$ as follows

$$E\{N(t)|x(t), z(t)\} = \mu_N(t)g(z(t)) \exp(x(t)'\beta)$$

and

$$\lambda_h(t|x(t), z(t)) = \lambda_{0h}(t)z(t) \exp(x(t)'\gamma),$$

where $\lambda_h(t)$ is the intensity function of $H(t)$, $x(t)$ is a vector of covariates that may depend on time, $z(t)$ is an unobserved stochastic process, and g is a completely unknown function.

In model (2), we have assumed that the observation process $H(t)$ is a Poisson process, which may not be true in practice. Further research is to replace model (2) by the following mean model

$$E\{H(t)|x, z\} = \mu_{0h}(t)z \exp(x'\gamma),$$

where $\mu_{0h}(t)$ is a completely unknown continuous baseline mean function. An estimation procedure needs to be developed for the joint mean models of the two counting processes.

Acknowledgements

The authors are grateful to the editor, Professor Jae Chang Lee, and two referees for their valuable comments and suggestions that greatly improved this article. The research of the first author was supported in part by a grant from the Research Grant Council of Hong Kong and a grant from the Hong Kong Polytechnic University. The research of the second author was supported in part by grants from the National Natural Science Foundation of China (No. 10971015) and the Key Project of Chinese Ministry of Education (No. 309007).

Appendix

Proof of Theorem 1. In this Appendix, we show the consistency and the asymptotic normality of $\hat{\beta}_{1n}$. Using the same notation as in Wang et al. (2001), define

$$Q(u) = \int_0^u G(v)\mathrm{d}\Lambda_{0h}(v),$$
$$R(u) = G(u)\Lambda_{0h}(u),$$
$$b_i(t) = \sum_{j=1}^{K_i^*}\left\{\int_t^\tau \frac{I(T_{ij} \le u \le C_i)\mathrm{d}Q(u)}{R^2(u)} - \frac{I(t \le T_{ij} \le \tau)}{R(T_{ij})}\right\},$$

and

$$f_{ih} = -E\left\{X_1K^*b_i(C)/\Lambda_{0h}(C)\right\} + x_{1i}\{K_i^*\Lambda_{0h}^{-1}(C_i) - \mathrm{e}^{x'_{1i}\gamma_*}\},$$

where the expectation is taken with respect to (X_1, C, K^*). Then we have

$$\widehat{\gamma}_n - \gamma_0 = n^{-1}\sum_{i=1}^n f_i + o_p(n^{-1/2}),$$

where f_i is the vector function $E(-\partial f_{ih}/\partial\gamma_*)^{-1}f_{ih}$ without the last entry (Wang et al., 2001; Huang and Wang, 2004).

The consistency of $\hat{\beta}_1$ follows from the two facts:

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

(ii)

$$\frac{\mathrm{d}}{\mathrm{d}\beta_1}U(\beta_1|\hat{\gamma}_n) = -\frac{1}{n}\sum_{i=1}^n x_{1i}x'_{1i}\mathrm{e}^{x'_{1i}\beta_1 + x'_i\hat{\gamma}_n}$$

converges uniformly to a negative matrix $-E(X'_1X_1\mathrm{e}^{X'_1\beta_1+X'\gamma_0})$ over β_1 in a neighborhood around the true value β_{10}.

Therefore the solution of the estimating function $\hat{\beta}_{1n}$ is unique and consistent. Now we turn to prove the asymptotical normality of the proposed estimator $\hat{\beta}_{1n}$.

Taylor expansion yields that

$$\begin{aligned}
U(\beta_1|\hat{\gamma}_n) &= \frac{1}{n}\sum_{i=1}^n x_{1i}(\bar{N}_i - \mathrm{e}^{x'_{1i}\beta_1+x'_i\hat{\gamma}_n}) \\
&= \frac{1}{n}\sum_{i=1}^n x_{1i}\left\{(\bar{N}_i - \mathrm{e}^{x'_{1i}\beta_1+x'_i\gamma_0}) - \mathrm{e}^{x'_{1i}\beta_1+x'_i\gamma_0}x'_i(\hat{\gamma}_n - \gamma_0)\right\} + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^n x_{1i}(\bar{N}_i - \mathrm{e}^{x'_{1i}\beta_1+x'_i\gamma_0}) - E(X_1\mathrm{e}^{X'_1\beta_1+X'\gamma_0}X')\frac{1}{n}\sum_{i=1}^n f_i + o_p(n^{-1/2}).
\end{aligned}$$

Define

$$U_i(\beta_1) = x_{1i}(\bar{N}_i - \mathrm{e}^{x'_{1i}\beta_1+x'_i\gamma_0}) - E(X_1\mathrm{e}^{X'_1\beta_1+X'\gamma_0}X')f_i,$$

then

$$U(\beta_1|\hat{\gamma}_n) = \frac{1}{n}\sum_{i=1}^n U_i(\beta_1) + o_p(n^{-1/2}).$$

By the standard procedure, one can obtain that the unique solution $\hat{\beta}_{1n}$ to $U(\beta_1|\hat{\gamma}_n) = 0$ satisfies the asymptotic normality. Specifically, $\sqrt{n}(\hat{\beta}_{1n} - \beta_{10})$ converges in distribution to a normal random variable with mean zero and the covariance matrix $\phi^{-1}\Sigma(\phi^{-1})'$, where $\phi = -E\left\{\left.\frac{\partial U_i(\beta_1)}{\partial\beta_1}\right|_{\beta_1=\beta_{10}}\right\}$ and $\Sigma = \mathrm{Cov}\{U_i(\beta_{10})\}$. This completes the proof. □

Proof of asymptotic properties of $\Phi_k(x)$ and $\Phi_0(t, x)$. In the following, we will only sketch the proof for the weak convergence of $\Phi_k(x)$ under models (1) and (2). The weak convergence of $\Phi_0(t, x)$ can be similarly derived. Assume that the limits of $S_k(t, x)$, $S_0(t)$, and $B(t, x)$ exist and are denoted by $s_k(t, x)$, $s_0(t)$, and $b(t, x)$, respectively. Define

$$M_i(t) = \int_0^t \left\{\Delta_i(u)H_i(u)\mathrm{d}N_i(u) - \Delta_i(u)\exp(\theta_0'X_i)\mathrm{d}\mathcal{A}(u)\right\}$$

where $\theta_0 = \beta_0 + \gamma_0$. To prove the weak convergence of $\Phi_k(x)$, first using Lemma A.1 of Lin and Ying (2001) and the functional version of the Taylor expansion, we have

$$\Phi_k(x) = n^{-1/2}\sum_{i=1}^n \int_0^\tau \left\{I(x_{ik} \le x) - \frac{s_k(u, x)}{s_0(u)}\right\}\mathrm{d}M_i(u) - b(\tau, x)'n^{1/2}(\hat{\theta} - \theta_0) + o_p(1).$$

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 $n^{1/2}(\hat{\theta} - \theta_0)$ converge in distribution and $b(\tau, x)$ are some deterministic functions. Thus $\Phi_k(x)$ is tight.

Let g_i be the vector $\phi^{-1}U_i(\beta_{10})$ without the last entry. Then, we can further write $\Phi_k(x)$ as

$$\Phi_k(x) = n^{-1/2}\sum_{i=1}^n \int_0^\tau \left\{I(x_{ik} \le x) - \frac{s_k(u, x)}{s_0(u)}\right\}\mathrm{d}M_i(u) - b(\tau, x)'n^{-1/2}\sum_{i=1}^n (f_i + g_i) + o_p(1).$$

It thus follows from the multivariate central limit theorem and the tightness of $\Phi_k(x)$ that $\Phi_k(x)$ converges weakly to a zero-mean Gaussian process which can be approximated by the zero-mean Gaussian process $\tilde{\Phi}_k(x)$ given in (5). □

References

Andrews, D.F., Herzberg, A.M., 1985. Data: A Collection of Problems from Many Fields for the Student and Research Worker. Springer-Verlag, New York.
Balakrishnan, N., Zhao, X., 2009. New multi-sample nonparametric tests for panel count data. Ann. Statist. 37, 1112–1149.
Balakrishnan, N., Zhao, X., 2010. A nonparametric test for the equality of counting processes with panel count data. Comput. Statist. Data Anal. 54, 135–142.
Byar, D.P., 1980. The veterans administration study of chemoprophylaxis for recurrent stage I bladder tumors: comparisons of placebo, pyridoxine, and topical thiotepa. In: Pavane-Macaluso, M., Smith, P.H., Edsmyr, F. (Eds.), Bladder Tumors and Other Topics in Urological Oncology. Plenum, New York, pp. 363–370.
Cheng, S.C., Wei, L.J., 2000. Inferences for a semiparametric model with panel data. Biometrika 87, 89–97.
He, X., Tong, X., Sun, J., 2009. Semiparametric analysis of panel count data with correlated observation and follow-up times. Lifetime Data Anal. 15, 177–196.
Hu, X.J., Sun, J., Wei, L.J., 2003. Regression parameter estimation from panel counts. Scand. J. Statist. 30, 25–43.
Huang, C.Y., Wang, M.C., 2004. Joint modeling and estimation for recurrent event processes and failure time data. J. Amer. Statist. Assoc. 99, 1153–1165.
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. J. Amer. Statist. Assoc. 80, 863–871.
Lin, D.Y., Wei, L.J., Yang, I., Ying, Z., 2000. Semiparametric regression for the mean and rate function of recurrent events. J. R. Stat. Soc. Ser. B 69, 711–730.
Lin, D.Y., Ying, Z., 2001. Semiparametric and nonparametric regression analysis of longitudinal data. J. Amer. Statist. Assoc. 96, 103–126.
Lipsitz, S.R., Fitzmaurice, G.M., Ibrahim, J.G., Gelber, R., Lipshultz, S., 2002. Parameter estimation in longitudinal studies with outcome-dependent follow-up. Biometrics 58, 621–630.
Lu, M., Zhang, Y., Huang, J., 2007. Estimation of the mean function with panel count data using monotone polynomial splines. Biometrika 94, 1–14.
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. Statist. 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. J. R. Stat. Soc. Ser. B 62, 293–302.
Thall, P.F., Lachin, J.M., 1988. Analysis of recurrent events: nonparametric methods for random-interval count data. J. Amer. Statist. Assoc. 83, 339–347.
Wang, M.C., Qin, J., Chiang, C.T., 2001. Analyzing recurrent event data with informative censoring. J. Amer. Statist. Assoc. 96, 1057–1065.
Wellner, J.A., Zhang, Y., 1998. Large sample theory for an estimator of the mean of a counting process with panel count data. Technical Report. Department of Statistics, University of Washington, Seattle.
Wellner, J.A., Zhang, Y., 2000. Two estimators of the mean of a counting process with panel count data. Ann. Statist. 28, 779–814.
Wellner, J.A., Zhang, Y., 2007. Two likelihood-based semiparametric estimation methods for panel count data with covariates. Ann. Statist. 35, 2106–2142.
Zhang, Y., 2002. A semiparametric pseudolikelihood estimation method for panel count data. Biometrika 89, 39–48.
Zhang, Y., 2006. Nonparametric k-sample tests for panel count data. Biometrika 93, 777–790.
Zhang, Y., Jamshidian, M., 2003. The gamma-frailty Poisson model for the nonparametric estimation of panel count data. Biometrics 59, 1099–1106.