PLOS ONE

RESEARCH ARTICLE

Quantile estimation of semiparametric model with time-varying coefficients for panel count data

Yijun Wang[1,2], **Weiwei Wang**[1,2]*

1 School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, Zhejiang Province, China, **2** Collaborative Innovation Center of Statistical Data Engineering, Technology & Application, Zhejiang Gongshang University, Hangzhou, Zhejiang Province, China

* weiweiwang@zjgsu.edu.cn

Citation: Wang Y, Wang W (2021) Quantile estimation of semiparametric model with time-varying coefficients for panel count data. PLoS ONE 16(12): e0261224. https://doi.org/10.1371/journal.pone.0261224

Editor: Feng Chen, Tongji University, CHINA

Received: June 7, 2021

Accepted: November 27, 2021

Published: December 13, 2021

Data Availability Statement: The data of bladder cancer data can be found in the Table 9.2 of the book "Statistical analysis of panel count data" (Sun and Zhao, 2013). Besides, the 2015 US flight delay data can be obtained from https://www.kaggle.com/usdot/flight-delays. The authors had no special access privileges to data that others would not have.

Funding: This paper was partially supported by the National Natural Science Foundation of China under Grand No. 12001485; the National Bureau of Statistics of China under Grand No. 2020LY073, and the Characteristic & Preponderant Discipline of

Abstract

Panel count data frequently occurs in follow-up studies, such as medical research, social sciences, reliability studies, and tumorigenicity experiences. This type data has been extensively studied by various statistical models with time-invariant regression coefficients. However, the assumption of invariant coefficients may be violated in some reality, and the temporal covariate effects would be of great interest in research studies. This motivates us to consider a more flexible time-varying coefficient model. For statistical inference of the unknown functions, the quantile regression approach based on the B-spline approximation is developed. Asymptotic results on the convergence of the estimators are provided. Some simulation studies are presented to assess the finite-sample performance of the estimators. Finally, two applications of bladder cancer data and US flight delay data are analyzed by the proposed method.

Introduction

In longitudinal follow-up studies, panel count data is frequently encountered in many fields such as medical research, social sciences, reliability studies, and tumorigenicity experiences, which has been widely analyzed by many authors. This type data is usually collected from the discrete observations in recurrent event process, as the continuous observations might be too expensive to be carried out. Thus, we can only obtain the cumulative occurrence numbers of the events of interest at these discrete observation times.

For the analysis of panel count data, [1, 2] developed the regression analysis approaches to the panel count data model. [3] studied the clustered mixed nonhomogeneous Poisson models of panel count data. [4] considered the spline-based likelihood estimation of the proportional mean model. To describe the potential correlations of the recurrent event process, [5–7] developed some joint models of panel count data by employing some frailty parameters to discuss these correlations. Recently, semiparametric transformation models with informative observation times were studied by many authors, such as [8–10]. More comprehensive introductions about this type data can be referred to the book of [11].

Key Construction Universities in Zhejiang Province (Zhejiang Gongshang University-Statistics). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

In general, the existing approaches in modeling panel count data are based on the time-invariant coefficients assumption, but which may be violated in practice. In some applications, coefficients may be time-varying, and sometimes it is more vital to detect the temporal impacts on the recurrent event process. For example, in medical studies, we are interested in detecting the temporal impacts of one new drug. Recently, [12, 13] proposed the varying coefficient models for recurrent events. However, the analysis of panel count data with varying coefficients is very limited. Most recently, [14] proposed a partially varying coefficient model of panel count data to account for the nonlinear interactions between covariates. [15] proposed a nonparametric proportional mean model of the panel count data with time-varying coefficients.

Quantile regression is widely used in the analysis of longitudinal data. It can provide more information about the distribution shape of the response and can be used to measure the effect of variables under different percentiles of the distribution. However, quantile regression methodologies for the panel count data are lagging. As the discreteness of the panel count data, quantile regression cannot be directly used. For the first, a smoothing technique ("jitter") is used for this type data, then the quantile regression can be applied to the smooth data.

In this paper, a semiparametric time-varying coefficient model is formulated. For the inference of the unknown functions, quantile regression method is used for the panel count data, with the unknown functions approximated by the B-spline basis functions. Furthermore, the asymptotic results on the convergence of the estimators are established as well. The main contribution of the paper is that we propose a new spline-based quantile estimation procedure for the time-varying coefficient panel count data model, which has not been discussed in the literature.

Model specification

Suppose that n independent subjects are observed over time. $N_i(t)$ denotes the cumulative total number of recurrent event occurring at or before time t for subject i. $\tilde{H}_i(t)$ is a counting process with jumps at the discrete observation times, $t_{i,1} < t_{i,2} < \cdots$. We assume that t is in a fix interval $\Re$ of finite length. Besides, two follow-up times are existed: the potential censoring time C_i^* and the observation endpoint T_i. Thus, only $C_i = \min(C_i^*, T_i)$ can be observed in the process, with $\delta_i = I(C_i = C_i^*)$. C_i^* is assumed to be independent with $N_i(t)$ and $\tilde{H}_i(t)$. Let $H_i(t) = \tilde{H}_i\{\min(t, C_i)\}$ denote the real observation process of subject i, and $m_i = \tilde{H}_i(C_i)$, $i = 1, \cdots, n$. Then, $N_i(t)$ can be only acquired at the time points where $H_i(t)$ jumps. The total number of the observations is defined as $m = \sum_{i=1}^{n} m_i$. Let Z_i be a $p \times 1$ vector of covariates. So we can have the independent and identically distributed dataset $\{H_i(t), N_i(t)dH_i(t), C_i, \delta_i, Z_i; t \geq 0, i = 1, \cdots, n\}$.

To describe the possible time-varying effects of covariates on $N_i(t)$, the time-varying coefficient model is proposed as follows.

- (1) Given Z_i, the conditional mean function of $N_i(t)$ is

$$E\{N_i(t)|Z_i\} = \int_0^t \lambda_0(u)\exp\{\beta(u)^\top Z_i\}du, \tag{1}$$

  where $\lambda_0(u)$ is an unspecified smooth baseline intensity function, and $\beta(u)$ is an unknown $p \times 1$ vector of time-varying regression coefficients.

- (2) Conditional on Z_i, $\{C_i, N_i(t), \tilde{H}_i(t)\}$ are mutually independent.

For the model defined above, [15] developed the likelihood and pseudo-likelihood methods to get the estimation of the baseline intensity function $\lambda_0(u)$ and the varying coefficient functions $\beta(u)$ based on the Poisson distribution assumption on $N_i(t)$. However, no distribution assumption is specified in this paper and the existed methods cannot be used. In the next section, the spline-based quantile regression is proposed to acquire the estimation of the unknown functions. In the first step, the unknown baseline intensity function and the coefficients are approximated by B-splines. And then, the discrete panel count data become continuous by a smoothing technique. Quantile regression is developed for the inference in the last step.

Estimation procedure

For the inference of Eq (1), the model can be rewritten as,

$$E\{N_i(t)|X_i\} = \int_0^t \exp\{X_i^\top \eta(u)\}du,$$

where $X_i = (Z_i^\top, 1)^\top$, $\eta(u) = (\beta(u)^\top, \log\{\lambda_0(u)\})^\top$.

Approximations of baseline and varying coefficients

Similar as [16], we use the basis expansion method to get the estimation of the unknown functions in this paper. Suppose $\eta_k(u)$, $k = 1, 2, \cdots, p+1$, can be approximated by a basis expansion, that is

$$\eta_k(u) \approx \sum_{l=1}^{L_k} \gamma_{kl} B_{kl}(u) = B_k(u)^\top \gamma_k,$$

where $B_k(u) = \{B_{k1}(u), \ldots, B_{kL_k}(u)\}^\top$ are basis functions, $\gamma_k = (\gamma_{k1}, \ldots, \gamma_{kL_k})^\top$ and L_k is the number of basis functions. Various basis functions can be used in the expansion such as Fourier basis functions, polynomial basis functions and B-spline functions. In this paper, the B-spline basis is selected in the estimation procedure for calculation simplicity.

The tuning parameter L_k is selected by $L_k = n_k + q_k + 1$, where n_k is the number of interior knots and q_k is the degree of the B-spline functions. The interior knots of the splines are equally spaced or placed on the sample quantiles of the data in all simulations and applications. The tuning parameter L_k may be different for different k. In this paper, we assume that $L_k = L$ and $q_k = q$ for all $\eta_k(u)$. Thus, we define $B_k(u) = B(u)$ for simplicity presentation.

Quantile regression

As quantile regression is a good alternative to the conditional mean models, the quantile regression is considered for the panel count data model. However, quantile regression cannot be directly used as the discreteness of the data $N_i(t)$. According to the method developed in [17], the "jitter" method is applied to construct continuous random variables. By adding U_{ij}, which is generated from a [0, 1) uniform distribution, we can have

$$N_i^*(t_{ij}) = N_i(t_{ij}) + U_{ij},$$

where the noise U_{ij} is independent of $N_i(t_{ij})$ and Z_i. The uniform distribution is used because it allows computational simplifications. The uniform noise, however, is by no means a necessity to jitter the data. The noise may be generated by any continuous distribution with support on [0, 1). Thus, we can get the continuous data $N_i^*(t_{ij})$ and there exists a one-to-one link between

the quantiles of $N_i(t_{ij})$ and $N_i^*(t_{ij})$. The regression model of N_i^* can be written as

$$N_i^*(t_{ij}) = \int_0^t \exp\{X_i^\top \eta(u)\} du + \epsilon_{ij},$$

where ϵ_{ij} are assumed to be independent of t_{ij} with unknown cumulative distribution function (cdf) $G(\cdot)$ and density function $g(\cdot)$. Besides, the τ-th conditional quantile of ϵ_{ij} is b_τ.

The quantile regression loss function is defined as $\rho_\tau(u) = u[\tau - I(u < 0)]$, $\tau \in (0, 1)$. Then the quantile regression is applied on the smooth data $N_i^*(t_{ij})$ to obtain the estimation of the unknown parameters. Thus, the unknown parameters $\phi = (\gamma^\top, b_\tau)^\top$ can be estimated by minimizing the following objective function $\Psi(\phi)$, that is

$$\Psi(\phi) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} \rho_\tau\{N_i(t_{ij}) - \int_0^{t_{ij}} \exp\{W(u, X_i)^\top \gamma\} du - b_\tau\},$$

where $W(u, X_i) = I_{p+1} \otimes B(u) \cdot X_i$ and $\gamma = (\gamma_1^\top, \ldots, \gamma_{p+1}^\top)^\top$.

For the ease of calculation, Gauss-Legendre formula is used to approximate the integral. Thus, we have

$$\int_0^{t_{ij}} \exp\{W(u, X_i)^\top \gamma\}\, du \approx \frac{t_{ij}}{2} \sum_{s=1}^{S} \omega_s \exp\left[W\left\{ \frac{t_{ij}}{2}(1 + \Delta_s), X_i \right\}^\top \gamma \right],$$

where ω_s is the Gauss coefficient, S is the number of the Gauss points and Δ_s is the Gauss point. The Gauss-Legendre approximation of the objective function $\Psi(\phi)$ can be defined as

$$\Psi(\phi) \approx \sum_{i=1}^{n}\sum_{j=1}^{m_i} \rho_\tau \left\{ N_i(t_{ij}) - \frac{t_{ij}}{2} \sum_{s=1}^{S} \omega_s \exp\left[W\left\{ \frac{t_{ij}}{2}(1 + \Delta_s), X_i \right\}^\top \gamma \right] - b_\tau \right\}.$$

Define $\hat{\phi} = (\hat{\gamma}^\top, \hat{b}_\tau)^\top$ be the minimizers of the approximation of the objective function $\Psi(\phi)$. It is nature to get the estimation of the varying coefficient $\beta_k(u)$, $k = 1, \cdots, p$,

$$\hat{\beta}_k(u) \approx \sum_{l=1}^{L} \hat{\gamma}_{kl} B_l(u) = B(u)^\top \hat{\gamma}_k,$$

and the baseline intensity function of $\lambda_0(u)$ can be obtained by

$$\hat{\lambda}_0(u) \approx \exp\left\{ \sum_{l=1}^{L} \hat{\gamma}_{p+1,l} B_l(u) \right\} = \exp\{B(u)^\top \hat{\gamma}_{p+1}\}.$$

Next, we discuss how to select the tuning parameter L and the Gauss point number S. As proposed by [16], we use the leave-one-subject-out cross-validation (CV) to choose L and S. Let $\hat{\gamma}^{(-i)}$ and $\hat{b}_\tau^{(-i)}$ denote the estimators from the data with the i-th subject deleted. So the leave-one-subject-out CV can be written as

$$CV(L, S) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} \rho_\tau \left\{ N_i^*(t_{ij}) - \frac{t_{ij}}{2} \sum_{s=1}^{S} \omega_s \exp\left[W\left\{ \frac{t_{ij}}{2}(1 + \Delta_s), X_i \right\}^\top \gamma^{(-i)} \right] - b_\tau^{(-i)} \right\}.$$

Thus, the tuning parameter L and S can be selected as

$$(L_{CV}, S_{CV}) = \min_{L,S} CV(L, S).$$

Remark 1 *The number L_k of the basis expansion of β_k may be different from each other. However, we assume $L_k = L$ for all k, for simplicity.*

Asymptotic results

The asymptotic results are concluded in this section. Before presenting the results, some regularity conditions are introduced for the first.

(C1) Z_i is uniformly bounded.

(C2) The observation number m_i is bounded by a constant.

(C3) $\lambda_0(u)$ and $\beta_k(u)$, $k = 1, \cdots, p$, are l-th differentiable and bounded.

(C4) There exists an open subset $\Omega \subset R^{pL+1}$, which contains the true parameter ϕ^*. The second derivative matrix $\nabla^2 h(t_{ij}, X_i;\phi)$ of $h(t_{ij}, X_i;\phi)$ with respect to ϕ, satisfies

$$\| \nabla^2 h(t_{ij}, X_i; \phi_1) - \nabla^2 h(t_{ij}, X_i; \phi_2) \| \leq M_1(t_{ij}, X_i) \| \phi_1 - \phi_2 \|,$$

$$\left| \frac{\partial^2 h(t_{ij}, X_i; \phi)}{\partial \phi_j \partial \phi_k} \right| \leq M_{2jk}(t_{ij}, Z_i),$$

for all $\phi \in \Omega$, with $E[M_1^2(t_{ij}, X_i)] < \infty$, $E[M_{2jk}^2(t_{ij}, X_i)] < \infty$ for all j, k.

(C5) $Var(\nabla h_{ij}^*) = M > 0$, $E\{(\nabla h_{ij}^*)^{\otimes 2}\} = \Gamma$, and $0 < d_1 < \lambda_{\min}(\Gamma) \leq \lambda_{\max}(\Gamma) < d_2 < \infty$, where $\lambda_{\min}(\Gamma)$ and $\lambda_{\max}(\Gamma)$ denote the smallest and the largest eigenvalues of Γ.

(C6) ϵ_{ij} is independent with unknown distribution function $G(\cdot)$ and density $g(\cdot)$. Besides, the τ-th conditional quantile of ϵ_{ij} is ℓ_τ.

Under these above regularity conditions, the asymptotic results on the convergence of the estimators are displayed in the following theory. For the need of the proofs, a lemma of spline function of [18] is presented. First, define

$$S_{kn} = \{\eta_{kn} : \eta_{kn} = \sum_{l=1}^{L} \gamma_{kl} B_l(u), (\gamma_{k1}, \cdots, \gamma_{kL}) \in R^L\}.$$

Let S_{kn} be the space of splines of degree q consisting of functions η_{kn} satisfying: (i) the function η_{kn} to each subinterval is a polynomial spline of degree q; (ii) for $q \geq 1$ and $0 \leq q' \leq q$, η_{kn} is q' times continuously differentiable on the support. Besides, η_k is assumed to satisfy the following regularity condition. Let $l_1 \in [0, q]$ be a nonnegative integer. The l_1-th derivative, denoted as $\eta_k^{(l_1)}$, exists and satisfies the Lipschitz condition of order $v \in (0, 1]$ such that $\rho = l_1 + v > 0.5$ and $|\eta_k^{(l_1)}(s) - \eta_k^{(l_1)}(t)| \leq \delta |s - t|^v$, for $s, t \in [0, C]$, where δ is a positive constant.

Lemma 1 *There exists $\eta_{kn} \in S_{kn}$ such that $\|\eta_{kn} - \eta_k\|_2 = O_p(L^{-\rho} + L^{1/2} m^{-1/2})$. If $L = O\{m^{1/(2\rho+1)}\}$, then we have $\|\eta_{kn} - \eta_k\|_2 = O_p\{(L/m)^{1/2}\} = O_p\{m^{-\rho/(2\rho+1)}\}$.*

Theorem 1 *Suppose the conditions (C1)–(C6) hold and if $L = O\{m^{1/(2\rho+1)}\}$, then we have*

$$\sqrt{m}(\hat{\phi} - \phi^*) \rightarrow_d N\{0, g(b_\tau^*)^{-2} \tau(1 - \tau)(\Gamma^{-1})\}.$$

Furthermore, we have

$$\| \hat{\beta}_k(u) - \beta_k(u) \|_2 = O_p\{(L/m)^{1/2}\}, k = 1, \cdots, p,$$
$$\| \log \hat{\lambda}_0(u) - \log \lambda_0(u) \|_2 = O_p\{(L/m)^{1/2}\}.$$

Ignoring the approximation error in the B-spline basis approximation of $\beta_k(u)$, $k = 1, \cdots, p$, we can have the $100(1-\alpha)\%$ pointwise confidence interval of $\beta_k(u)$ under quantile τ,

$$\hat{\beta}_k(u) \pm z_{2/\alpha}\sqrt{cov\{\hat{\beta}_k(u)\}},$$

where $z_{2/\alpha}$ is the $100(1-\alpha)\%$ percentile of the standard normal distribution and $cov\{\hat{\beta}_k(u)\} = B(u)^\top cov(\hat{\gamma}_k)B(u)$. Similar as the baseline function $\lambda_0(u)$.

Simulation studies

Three simulation studies are carried out to evaluate the performance of the method developed in this paper. We generated 200 datasets from the time-varying coefficient model, each of size $n = 100$ or 200 independent subjects. For each subject i, the endpoint of observation T_i is assumed to be 6 and the censoring time C_i^* follows the uniform distribution of $[T_i/2, 3T_i/2]$. The number of observation times m_i is generated from a discrete uniform distribution {1, 2, 3, 4, 5}. And the observed event times, $\{t_{i1}, \ldots, t_{im_i}\}$, are the order statistics of a random sample size m_i from the uniform distribution over $(0, C_i)$. Given m_i and $\{t_{i1}, \ldots, t_{im_i}\}$, the panel count data $N_i(t_{ij})$ can be obtained by the following formula

$$\begin{aligned} N_i(t_{ij}) &= N_i^*[\lambda_N(t_{i1})] + N_i^*[\lambda_N(t_{i2}) - \lambda_N(t_{i1})] \\ &\quad + \cdots + N_i^*[\lambda_N(t_{ij}) - \lambda_N(t_{ij-1})], \end{aligned}$$

for $j = 1, \cdots m_i$ and $i = 1, \cdots n$. $N_i^*[\lambda_N(t_{ij})]$ is the random number generated from the Poisson distribution with mean

$$\int_0^{t_{ij}} \lambda_0(u)\exp\{\beta(u)^\top Z_i\}du.$$

The following three cases are considered:

- Case I: $p = 1$ and the covariate Z_i is generated independently from the [0, 1] uniform distribution. The baseline function is taken as $\lambda_0(u) = 2u + 1$ and the varying coefficient $\beta(u) = \sin(-\pi u/6)$.
- Case II: $p = 1$ and the covariate Z_i is generated independently from the [0, 1] uniform distribution. The baseline function is taken as $\lambda_0(u) = 2(u + \tau)$ and the varying coefficient $\beta(u) = \sin(-\tau\pi u/6)$.
- Case III: $p = 2$ and the covariates Z_i are generated from the $[0, 1]^2$ uniform distribution with correlation $cor(Z_{ik}, Z_{il}) = 0.5^{|k-l|}$. The baseline function is taken as $\lambda_0(u) = 2u + 1$ and the varying coefficient $\beta_1(u) = sin(-\pi u/6)$ and $\beta_2(u) = 2sin(-\tau\pi u/6)$.

To estimate the smooth functions $\log\lambda_0(u)$ and $\beta(u)$, the cubic B-spline functions are selected. Under different quantiles $\tau = \{0.25, 0.5, 0.75\}$, the estimations of Case I–III are presented with sample size $n = 100$ or 200 in Tables 1–3, respectively. The results include the average of the absolute bias values based 100 grid points (BIAS), the average of sampling standard errors based 100 grid points (SSE), the average of the bootstrap standard errors based 100 grid

Table 1. BIAS, SSE, BSE and CP of the estimated functions in Case I at different τ.

τ	n	Estimated function	BIAS	SSE	BSE	CP
0.25	100	$\beta(t)$	0.0558	0.8329	0.8355	0.9640
		$\log \lambda_0(t)$	0.1276	0.4150	0.4453	0.9548
	200	$\beta(t)$	0.0319	0.6826	0.5999	0.9355
		$\log \lambda_0(t)$	0.1044	0.3853	0.3377	0.9470
0.5	100	$\beta(t)$	0.0598	0.4472	0.4184	0.9623
		$\log \lambda_0(t)$	0.0341	0.2460	0.2264	0.9750
	200	$\beta(t)$	0.0197	0.3905	0.3684	0.9611
		$\log \lambda_0(t)$	0.0280	0.1886	0.1725	0.9525
0.75	100	$\beta(t)$	0.0789	0.8014	0.9118	0.9701
		$\log \lambda_0(t)$	0.0751	0.4935	0.4955	0.9640
	200	$\beta(t)$	0.0553	0.6901	0.6785	0.9368
		$\log \lambda_0(t)$	0.0599	0.4079	0.3808	0.9472

https://doi.org/10.1371/journal.pone.0261224.t001

points (BSE) and the average of the estimated 95% coverage probabilities based 100 grid points (CP). It can be seen that the estimations are unbiased under different quantiles. The values of SSE and BSE are close and decrease with the increasing sample size n. Besides, from the results of CP, we can note that the Gaussian approximation is appropriate for the estimators.

Figs 1–3 display the estimation curves of the unknown functions $\log \lambda_0(t)$ and $\beta(t)$ with $n = 200$. In the figures, the point lines represent the estimated curves, the solid lines represent the true curves and the dotted lines represent the 95% confidence intervals. Based the figures, it is easy to find that the real curves and the estimated curves are very close, which indicates the B-spline estimations of the unknown functions work well. From the simulation results, we note that the estimations under different quantiles are reasonable for $\log \lambda_0(t)$ and $\beta(t)$.

Applications

Bladder cancer data

Bladder cancer data was collected by the Veterans Administration Cooperative Urological Research Group. In this study, 85 patients were randomly assigned to two treatment groups: placebo group (47) and thiotepa group (38). For each patient, the observation times and the cumulative numbers of the bladder tumors that occurring at or before the observation times

Table 2. BIAS, SSE, BSE and CP of the estimated functions in Case II at different τ.

τ	n	Estimated function	BIAS	SSE	BSE	CP
0.25	100	$\beta(t)$	0.0768	0.9830	0.9890	0.9592
		$\log \lambda_0(t)$	0.1501	0.4924	0.5356	0.9661
	200	$\beta(t)$	0.0214	0.7345	0.6980	0.9365
		$\log \lambda_0(t)$	0.1099	0.4021	0.3875	0.9480
0.5	100	$\beta(t)$	0.0455	0.7163	0.6937	0.9250
		$\log \lambda_0(t)$	0.0746	0.4082	0.4277	0.9425
	200	$\beta(t)$	0.0380	0.4600	0.4511	0.9389
		$\log \lambda_0(t)$	0.0514	0.2633	0.2601	0.9694
0.75	100	$\beta(t)$	0.0552	0.8944	0.8558	0.9697
		$\log \lambda_0(t)$	0.0722	0.5191	0.4701	0.9406
	200	$\beta(t)$	0.0398	0.6906	0.6721	0.9355
		$\log \lambda_0(t)$	0.0529	0.3553	0.3284	0.9640

https://doi.org/10.1371/journal.pone.0261224.t002

Table 3. BIAS, SSE, BSE and CP of the estimated functions in Case III at different τ.

τ	n	Estimated function	BIAS	SSE	BSE	CP
0.25	100	$\beta_1(t)$	0.0553	1.0039	0.9841	0.9560
		$\beta_2(t)$	0.1272	0.9849	0.9762	0.9262
		$\log \lambda_0(t)$	0.1694	0.7364	0.6911	0.9735
	200	$\beta_1(t)$	0.0404	0.7205	0.6835	0.9436
		$\beta_2(t)$	0.0795	0.7557	0.7422	0.9677
		$\log \lambda_0(t)$	0.1271	0.5135	0.4974	0.9595
0.5	100	$\beta_1(t)$	0.1330	0.8475	0.8618	0.9205
		$\beta_2(t)$	0.1641	0.9111	0.8852	0.9455
		$\log \lambda_0(t)$	0.0431	0.5986	0.5748	0.9625
	200	$\beta_1(t)$	0.0927	0.6934	0.6520	0.9380
		$\beta_2(t)$	0.0518	0.7482	0.7249	0.9628
		$\log \lambda_0(t)$	0.0445	0.4225	0.4418	0.9561
0.75	100	$\beta_1(t)$	0.1235	0.9436	0.8869	0.9256
		$\beta_2(t)$	0.1282	0.9478	0.9648	0.9460
		$\log \lambda_0(t)$	0.0695	0.8331	0.8160	0.9335
	200	$\beta_1(t)$	0.0383	0.7438	0.7409	0.9510
		$\beta_2(t)$	0.0825	0.8510	0.8325	0.9246
		$\log \lambda_0(t)$	0.0507	0.5684	0.5348	0.9714

https://doi.org/10.1371/journal.pone.0261224.t003

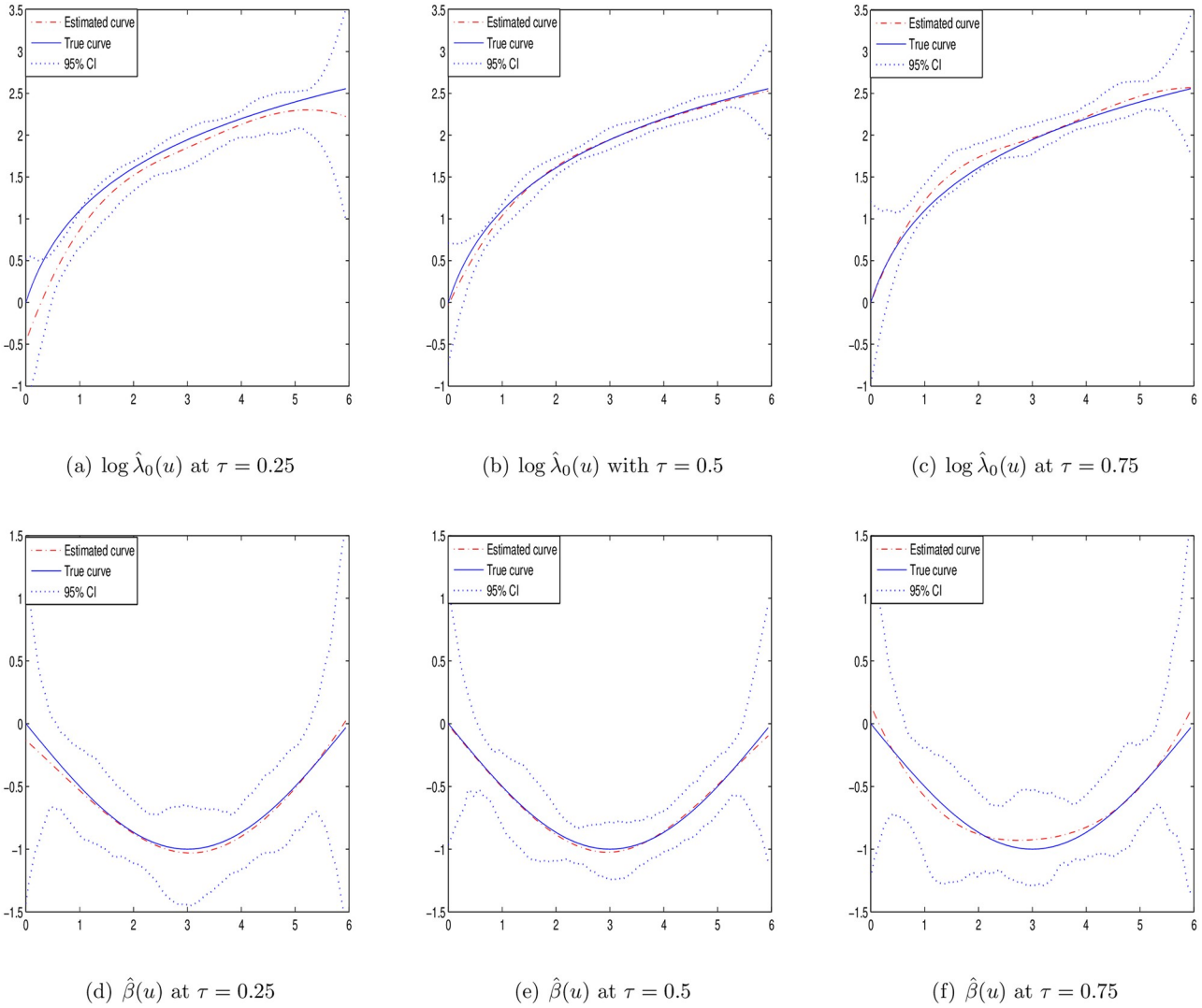
are recorded. The observation endpoint is 53 month. What's more, the initial number of the bladder tumors and the largest initial tumor size for each patient are also recorded. In the literature, the dataset has been discussed by many authors such as [5, 7, 19]. However, time-varying coefficient panel count data model is not considered for this dataset.

In order to describe the temporal impacts of the covariates on the bladder cancer data, the time-varying coefficient model proposed in this paper is applied to these data. For each patient i, $N_i(t)$ is denoted as the cumulative bladder tumors number occurring up to time t, and $H_i(t)$ is denoted as the cumulative observation number up to time t, $i = 1, \cdots, 85$. Furthermore, let $Z_{i1} = 1$ if the patient i is belonged to the thiotepa group and $Z_{i1} = 0$ otherwise. Z_{i2} is denoted as the initial tumor number and Z_{i3} is the natural logarithm of the largest initial tumour size plus 1 for each patient i. Therefore, we have the model

$$E\{N_i(t)|Z_i\} = \int_0^t \lambda_0(u) \exp\{\beta_1(u)Z_{i1} + \beta_2(u)Z_{i2} + \beta_3(u)Z_{i3}\}du.$$

Then quantile regression estimation is applied to this data. 100 samples are drawn from the data every time and 200 times are repeated in the estimation. Similar to the numerical studies, the unknown functions $\lambda_0(t)$ and $\beta_k(t)$, $k = 1, 2, 3$ are approximated by Cubic B-spline functions. The estimation is implemented under quantiles $\tau \in \{0.25, 0.5, 0.75\}$.

The estimation curves of $\log \lambda_0(t)$ and $\beta_k(t)$, $k = 1, 2, 3$ are displayed in Fig 4. In general, the thiotepa treatment and the tumor recurrence rate are negatively correlated at different quantiles. Patients in the thiotepa group tend to have less tumor recurrence rate than those in the placebo group. The initial tumor number is positively correlated with the recurrence rate and the largest initial tumor size is negatively correlated with the recurrence rate. These above conclusions are consistent with [19]. Furthermore, we can find the covariates impacts are varying during the observation time and the impacts are different at different quantiles. Thus, more information can be obtained from the quantile regression of the time-varying coefficient panel count data model than the other analysis in the existing literature.



(a) $\log \hat{\lambda}_0(u)$ at $\tau = 0.25$ (b) $\log \hat{\lambda}_0(u)$ with $\tau = 0.5$ (c) $\log \hat{\lambda}_0(u)$ at $\tau = 0.75$

(d) $\hat{\beta}(u)$ at $\tau = 0.25$ (e) $\hat{\beta}(u)$ at $\tau = 0.5$ (f) $\hat{\beta}(u)$ at $\tau = 0.75$

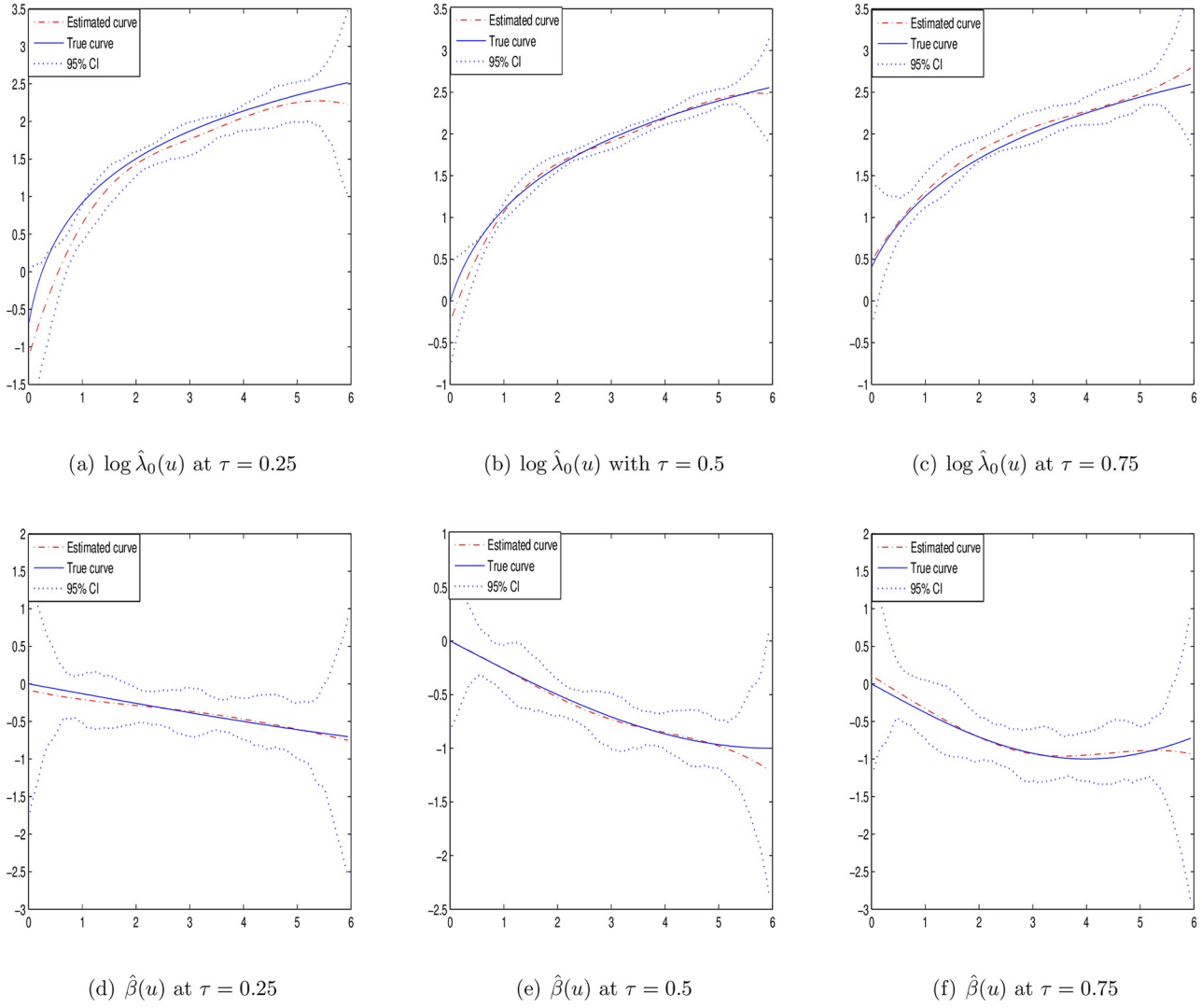
Fig 1. Estimated curves of time-varying functions in case I at different τ with $n = 200$.

https://doi.org/10.1371/journal.pone.0261224.g001

US flight delay data

In this subsection, 2015 US flight delay data (available from https://www.kaggle.com/usdot/flight-delays) is analyzed with the time-varying coefficient panel count data model. This dataset was collected from the U.S. Department of Transportation's (DOT) monthly Air Travel Consumer Report. The report contained information about the numbers of on-time, delayed, canceled, and diverted flights. The dataset included 9794 flights which were observed during 3 months in the year of 2015. The numbers of delays for each flight are recorded between the observation times. The observation times of each flight are the same and the observation interval is 7 days. Besides, the average departure delay time and the average flight distance of each flight are also recorded.

In order to describe the temporal covariates impacts on the flight delays, the time-varying coefficient model proposed in this paper is used to these data. For each flight i, $N_i(t)$ is denoted as the cumulative flight delay number that had occurred up to time t, $H_i(t)$ is denoted as the



(a) $\log \hat{\lambda}_0(u)$ at $\tau = 0.25$ (b) $\log \hat{\lambda}_0(u)$ with $\tau = 0.5$ (c) $\log \hat{\lambda}_0(u)$ at $\tau = 0.75$

(d) $\hat{\beta}(u)$ at $\tau = 0.25$ (e) $\hat{\beta}(u)$ at $\tau = 0.5$ (f) $\hat{\beta}(u)$ at $\tau = 0.75$

Fig 2. Estimated curves of time-varying functions in case II at different τ with $n = 200$.

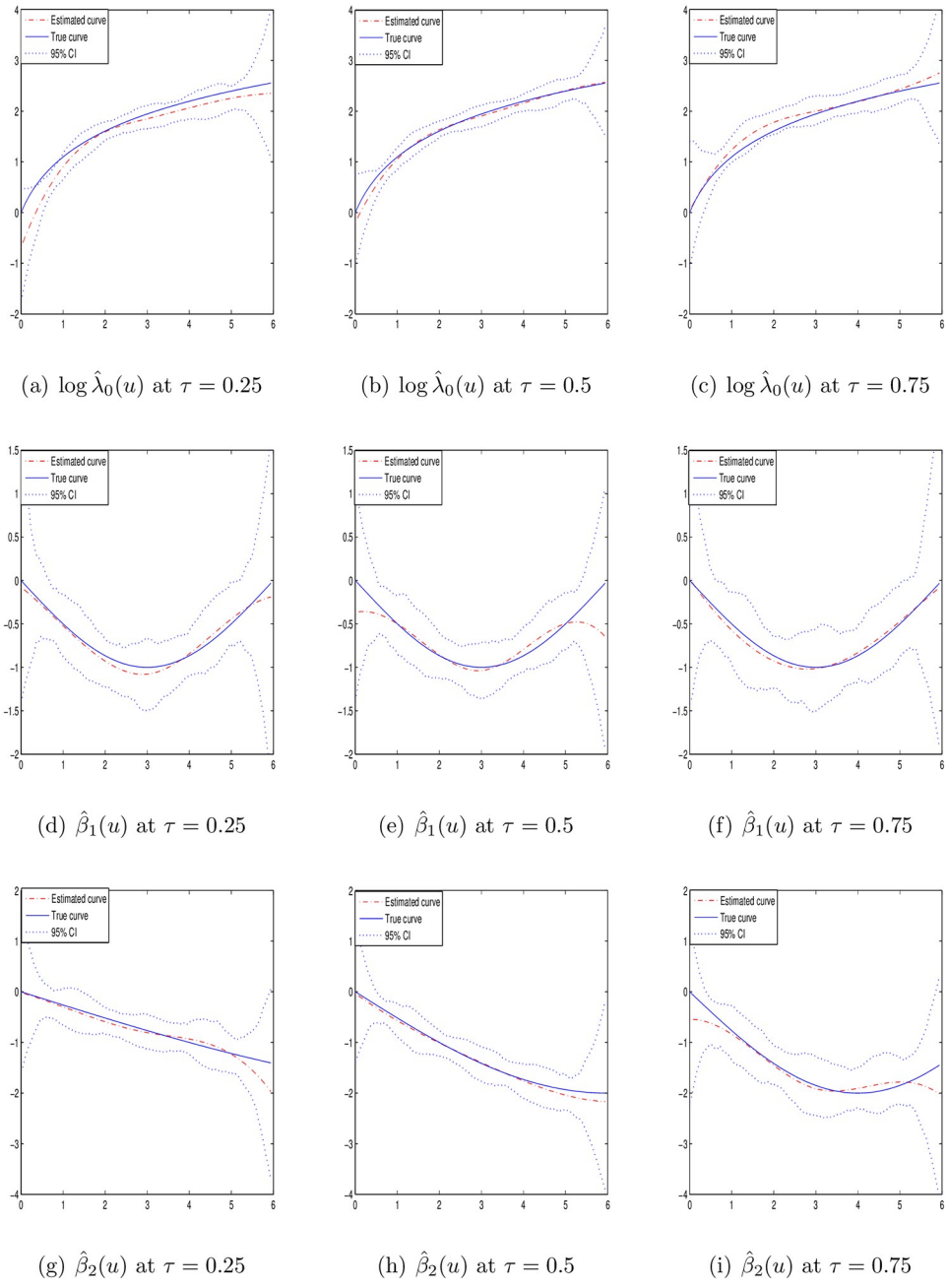
https://doi.org/10.1371/journal.pone.0261224.g002

cumulative observation number up to time t, $i = 1, \cdots, 9794$. Furthermore, we define Z_{i1} as the average time of the departure delay and Z_{i2} as the average distance of the flight i. Therefore, we have the model

$$E\{N_i(t)|Z_i\} = \int_0^t \lambda_0(u) \exp\{\beta_1(u)Z_{i1} + \beta_2(u)Z_{i2}\}du.$$

Then spline-based quantile estimation is applied to this data. Similarly, the unknown functions $\lambda_0(t)$ and $\beta_k(t)$, $k = 1, 2$ are also approximated by Cubic B-spline functions. The estimation is implemented under quantiles $\tau \in \{0.25, 0.5, 0.75\}$.

As the sample size of the dataset is large, it is time-consuming or even not possible to read the entire dataset in practice due to the limited memory. Besides, the direct analysis can be infeasible, mainly due to the computing memory or computing time. In order to deal with the massive data, parallel computing method is developed by [20, 21]. In parallel computing method, we split the original dataset into a family of disjoint sub-sample blocks with equal size



(a) $\log \hat{\lambda}_0(u)$ at $\tau = 0.25$ (b) $\log \hat{\lambda}_0(u)$ at $\tau = 0.5$ (c) $\log \hat{\lambda}_0(u)$ at $\tau = 0.75$

(d) $\hat{\beta}_1(u)$ at $\tau = 0.25$ (e) $\hat{\beta}_1(u)$ at $\tau = 0.5$ (f) $\hat{\beta}_1(u)$ at $\tau = 0.75$

(g) $\hat{\beta}_2(u)$ at $\tau = 0.25$ (h) $\hat{\beta}_2(u)$ at $\tau = 0.5$ (i) $\hat{\beta}_2(u)$ at $\tau = 0.75$

Fig 3. Estimated curves of time-varying functions in case III at different τ with $n = 200$.

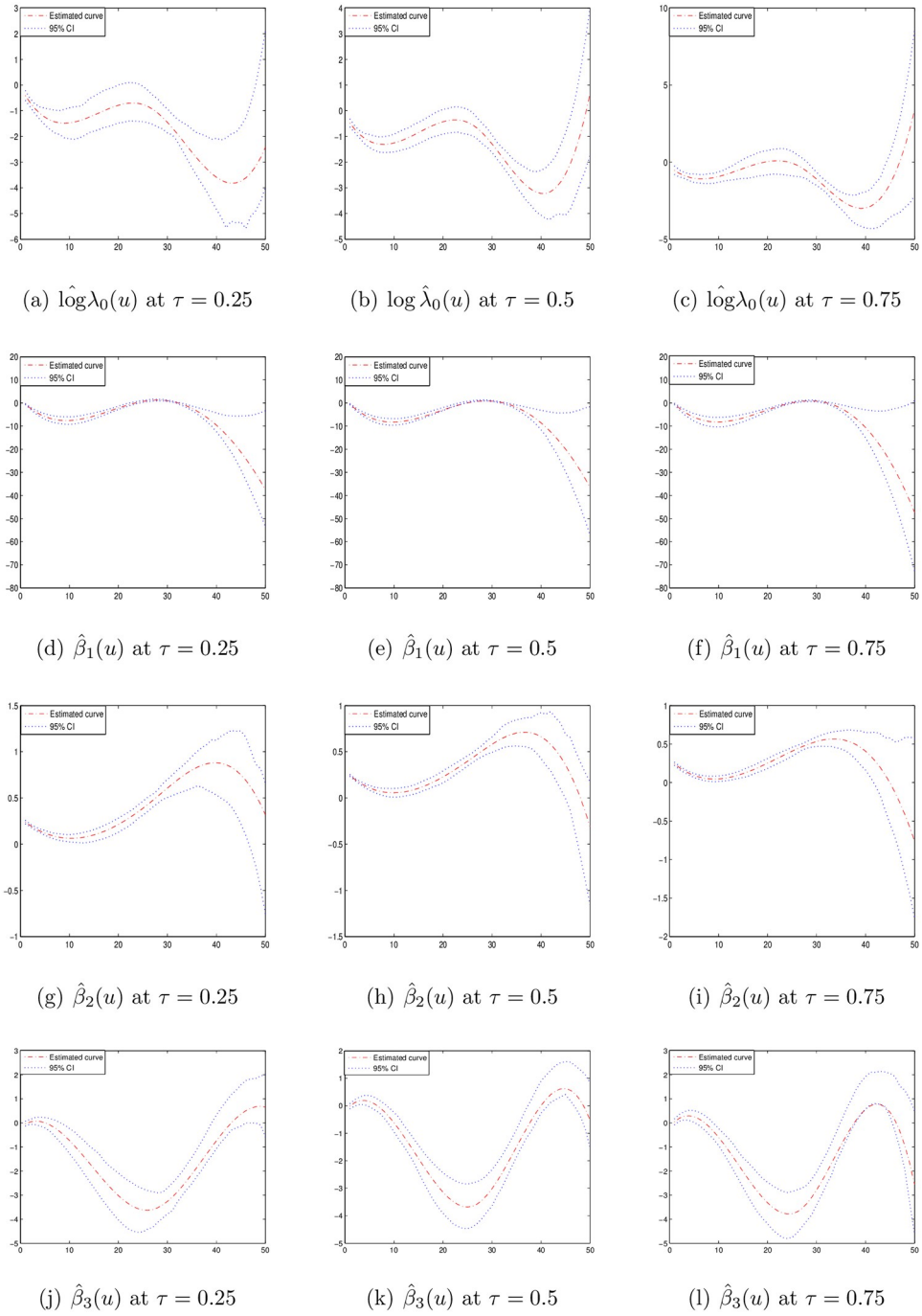
https://doi.org/10.1371/journal.pone.0261224.g003

for the first. More precisely, the data structure can be defined as the following form:

$$S = [S_k = \{H_{ki}(t), N_{ki}(t)dH_{ki}(t), Z_{ki}; t \geq 0, i = 1, \cdots, m\}, k = 1, \cdots, K],$$

where the original dataset S is of size $n = K \times m$ which is partitioned to K subsample blocks S_k each consist m samples which are randomly picked up from the dataset S.

Thus, the estimation procedure proposed can be implemented for every block S_k, $k = 1, \cdots, K$ and the estimated values of unknown parameters for each block S_k is denoted as $\{\hat{\eta}(t)\}_k$.



(a) $\widehat{\log}\lambda_0(u)$ at $\tau = 0.25$ (b) $\log\hat{\lambda}_0(u)$ at $\tau = 0.5$ (c) $\widehat{\log}\lambda_0(u)$ at $\tau = 0.75$

(d) $\hat{\beta}_1(u)$ at $\tau = 0.25$ (e) $\hat{\beta}_1(u)$ at $\tau = 0.5$ (f) $\hat{\beta}_1(u)$ at $\tau = 0.75$

(g) $\hat{\beta}_2(u)$ at $\tau = 0.25$ (h) $\hat{\beta}_2(u)$ at $\tau = 0.5$ (i) $\hat{\beta}_2(u)$ at $\tau = 0.75$

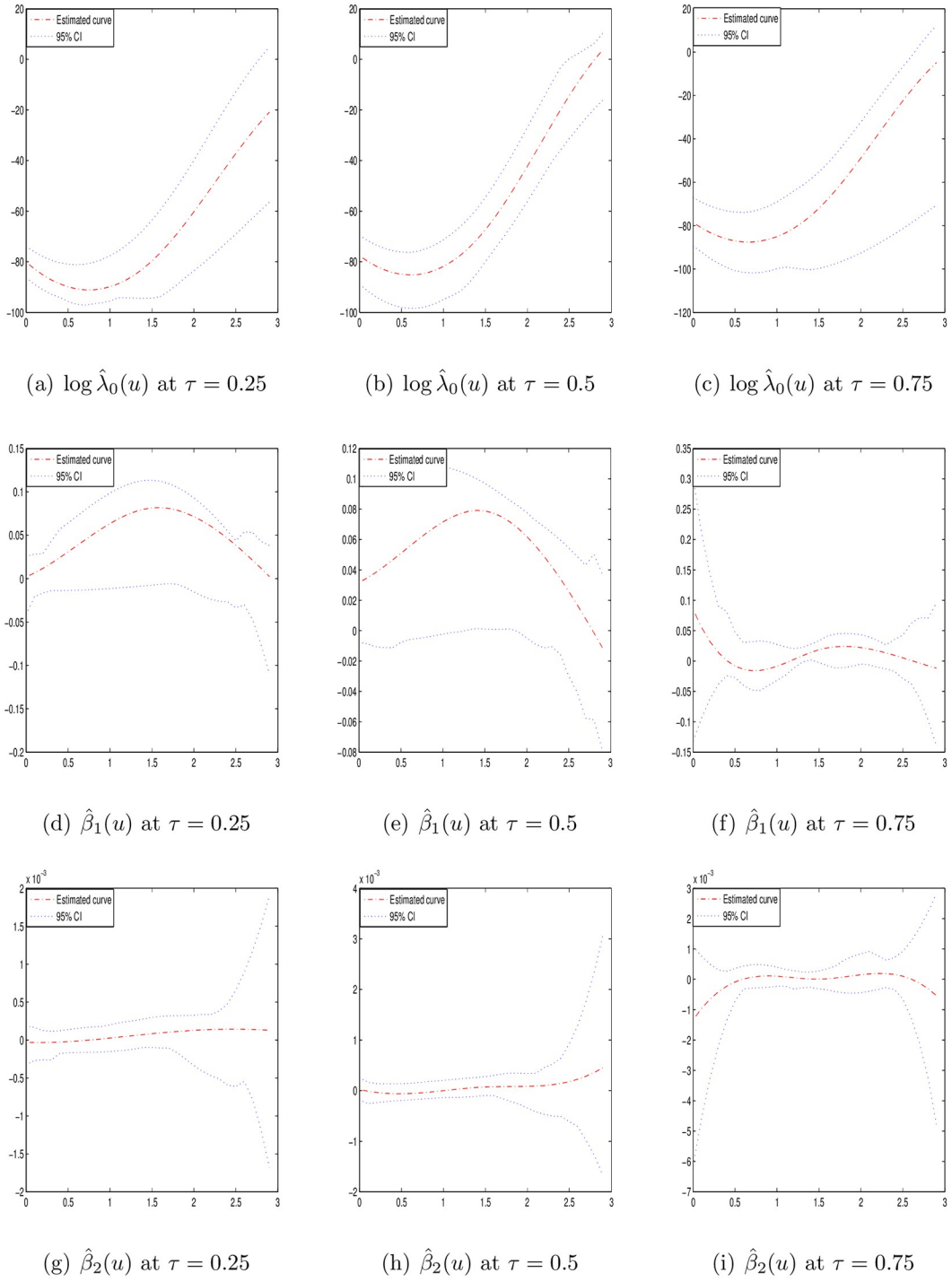
(j) $\hat{\beta}_3(u)$ at $\tau = 0.25$ (k) $\hat{\beta}_3(u)$ at $\tau = 0.5$ (l) $\hat{\beta}_3(u)$ at $\tau = 0.75$

Fig 4. Estimated curves of time-varying functions for bladder cancer data at different τ.

https://doi.org/10.1371/journal.pone.0261224.g004

Similar to the method introduced in [21], the final full-sample estimators can be generated by

$$\hat{\eta}(t) = \sum_{k=1}^{K} \{\hat{\eta}(t)\}_k.$$



(a) $\log \hat{\lambda}_0(u)$ at $\tau = 0.25$ (b) $\log \hat{\lambda}_0(u)$ at $\tau = 0.5$ (c) $\log \hat{\lambda}_0(u)$ at $\tau = 0.75$

(d) $\hat{\beta}_1(u)$ at $\tau = 0.25$ (e) $\hat{\beta}_1(u)$ at $\tau = 0.5$ (f) $\hat{\beta}_1(u)$ at $\tau = 0.75$

(g) $\hat{\beta}_2(u)$ at $\tau = 0.25$ (h) $\hat{\beta}_2(u)$ at $\tau = 0.5$ (i) $\hat{\beta}_2(u)$ at $\tau = 0.75$

Fig 5. Estimated curves of time-varying functions for US flight delay data at different τ.

https://doi.org/10.1371/journal.pone.0261224.g005

The estimation curves of $\log \lambda_0(t)$ and $\beta_k(t)$, $k = 1, 2$, under different quantiles $\tau \in \{0.25, 0.5, 0.75\}$ are displayed in Fig 5. From Fig 5, we can find that the departure delay time is positively correlated with the cumulative flight delay numbers. Besides, the impact of the departure delay time is varying over the time under different quantiles and the impact is different at different quantiles. However, the effect of the flight distance is not significant on the flight delay numbers.

Concluding remarks

In this paper, we proposed a spline-based quantile regression estimation method in the time-varying coefficient panel count data model. This model discussed in our paper is more general than [15], with no Poisson restriction on the recurrent event process. To get the estimations, B-splines are used to approximate the unknown functions $\log \lambda_0(t)$ and $\beta(t)$ for the first, and then a smoothing technique is applied to obtain the continuation of the discrete panel count data. Finally, the spline-based quantile regression approach is developed at different quantiles. Some simulations are presented to evaluate the performance of the proposed approach and two applications are analyzed to demonstrate its effectiveness in this paper.

Recently, the Enron e-mail corpus which was a massive set of the e-mail messages, have been discussed by many authors, such as [22]. If we are interested in the number of interactions of all pairs of individuals in this longitudinal observations, as usual in network analysis, the snapshots are applied to model this longitudinal networks, then, this is a standard panel count dataset with massive observations. Furthermore, in this paper, we only considered the situation with low dimensional covariates, which may be not unpracticable in the applications. As the high-dimensional covariates may be existed, variable selection methods can be considered for the time-varying coefficient model. This will be an important topic for our further studies. Besides, reliability data and traffic data have been studied by many authors, such as [23–26]. This will be interesting to study the quantile regression estimation of such data.

Proof of Theorem 1

Define γ^*, b_τ^* as the true but unknown values of γ, b_τ^*, $u_1 = \sqrt{m}(\gamma - \gamma^*)$, $u_2 = \sqrt{m}(b_\tau - b_\tau^*)$, $u = (u_1^\top, u_2)^\top$, $\phi = (\gamma^\top, b_\tau)^\top$ and $h(t_{ij}, X_i; \phi) = \int_0^{t_{ij}} \exp\{W(u, X_i)^\top \gamma\} du + b_\tau$.

Let

$$\begin{aligned} H(\phi^*) &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\rho_\tau\left\{N_i^*(t_{ij}) - \int_0^{t_{ij}} \exp\{W(u, X_i)^\top \gamma^*\} du - b_\tau^*\right\} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\rho_\tau\left\{\epsilon_{ij} - b_\tau^* + \int_0^{t_{ij}} \exp\{X_i^\top \eta(u)\} du - \int_0^{t_{ij}} \exp\{W(u, X_i)^\top \gamma^*\} du\right\} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\rho_\tau(\epsilon_{ij} - b_\tau^* + r_{ij}), \end{aligned}$$

where

$$r_{ij} = \int_0^{t_{ij}} \exp\{X_i^\top \eta(u)\} du - \int_0^{t_{ij}} \exp\{W(u, X_i)^\top \gamma^*\} du.$$

By the Taylor expansion, we can have $r_{ij} = o_p(1)$. Besides,

$$\begin{aligned} H(\phi^* + u/\sqrt{m}) &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\rho_\tau\{N_i^*(t_{ij}) - h(t_{ij}, X_i; \phi^* + u/\sqrt{m})\} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\rho_\tau\{\epsilon_{ij} - b_\tau^* + r_{ij} - \nabla h(t_{ij}, X_i; \tilde{\phi})^\top u/\sqrt{m}\} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\rho_\tau\{\epsilon_{ij} - b_\tau^* + r_{ij} - \zeta_{ij}\}, \end{aligned}$$

where $\zeta_{ij} = \nabla h(t_{ij}, X_i; \tilde{\phi})^\top u/\sqrt{m}$ and $\tilde{\phi}$ is between ϕ^* and $\phi^* + u/\sqrt{m}$. Define

$$\begin{aligned}\Delta H &= H(\phi^* + u/\sqrt{m}) - H(\phi^*) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\rho_\tau(\epsilon_{ij} - b_\tau^* + r_{ij} - \zeta_{ij}) - \sum_{i=1}^{n}\sum_{j=1}^{m_i}\rho_\tau(\epsilon_{ij} - b_\tau^* + r_{ij}).\end{aligned}$$

By the identity of [27],

$$|r - s| - |r| = -s\{I(r > 0) - I(r < 0)\} + 2\int_0^s \{I(r \le x) - I(r \le 0)\}dx.$$

Hence, it can be obtained that

$$\rho_\tau(r - s) - \rho_\tau(r) = s\{I(r < 0) - \tau\} + \int_0^s \{I(r \le x) - I(r \le 0)\}dx.$$

Thus, ΔH can be denoted as $\Delta H = \Delta H_1 + \Delta H_2$, with

$$\begin{aligned}\Delta H_1 &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\zeta_{ij}\{I(\epsilon_{ij} < b_\tau^* + r_{ij}) - \tau\}, \\ \Delta H_2 &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\int_0^{\zeta_{ij}}\{I(\epsilon_{ij} \le x + r_{ij} + b_\tau^*) - I(\epsilon_{ij} \le r_{ij} + b_\tau^*)\}dx.\end{aligned}$$

By calculating the expectation and variance of ΔH_2,

$$\begin{aligned}E(\Delta H_2) &= E\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i}\int_0^{\zeta_{ij}}\{I(\epsilon_{ij} \le x + r_{ij} + b_\tau^*) - I(\epsilon_{ij} \le r_{ij} + b_\tau^*)\}dx\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}E\left[\int_0^{\zeta_{ij}}\{I(\epsilon_{ij} \le x + r_{ij} + b_\tau^*) - I(\epsilon_{ij} \le r_{ij} + b_\tau^*)\}dx\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\int_0^{\zeta_{ij}}\{G(x + r_{ij} + b_\tau^*) - G(r_{ij} + b_\tau^*)\}dx \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\int_0^{\zeta_{ij}}\{G(x + r_{ij} + b_\tau^*) - G(r_{ij} + b_\tau^*)\}dx \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\int_0^{\zeta_{ij}} xg(r_{ij} + b_\tau^*)dx + o_p(1) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{\zeta_{ij}^2}{2}g(r_{ij} + b_\tau^*) + o_p(1) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{\zeta_{ij}^2}{2}g(b_\tau^*) + o_p(1).\end{aligned}$$

By condition (C5), $E[\{\nabla h(t_{ij}, X_i;\phi^*)\}^{\otimes 2}] = \Gamma$, we can have

$$\begin{aligned} E(\Delta H_2) &= \frac{g(b_\tau^*)}{2m}\sum_{i=1}^{n}\sum_{j=1}^{m_i} u^\top \nabla h(t_{ij}, X_i;\tilde{\phi})\nabla h(t_{ij}, X_i;\tilde{\phi})^\top u \\ &= \frac{g(b_\tau^*)}{2} u^\top \sum_{i=1}^{n}\sum_{j=1}^{m_i} \frac{\nabla h(t_{ij}, X_i;\phi^*)\nabla h(t_{ij}, X_i;\phi^*)^\top}{m} u + o_p(1) \\ &= \frac{g(b_\tau^*)}{2} u^\top \Gamma u + o_p(1). \end{aligned}$$

Next, we calculate the variance of ΔH_2,

$$\begin{aligned} Var(\Delta H_2) &= Var\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i}\int_0^{\zeta_{ij}} \{I(\varepsilon_{ij} \le x + r_{ij} + b_\tau^*) - I(\epsilon_{ij} \le r_{ij} + b_\tau^*)\}dx\right] \\ &\le \sum_{i=1}^{n}\sum_{j=1}^{m_i} E\left[\int_0^{\zeta_{ij}} \{I(\epsilon_{ij} \le x + r_{ij} + b_\tau^*) - I(\epsilon_{ij} \le r_{ij} + b_\tau^*)\}dx\right]^2 \\ &\le \sum_{i=1}^{n}\sum_{j=1}^{m_i}\int_0^{|\zeta_{ij}|}\int_0^{|\zeta_{ij}|} \{G(|\zeta_{ij}| + r_{ij} + b_\tau^*) - G(r_{ij} + b_\tau^*)\}dx_1 dx_2 = o_p(1). \end{aligned}$$

Hence, we can have $\Delta H_2 = (1/2)g(b_\tau^*)u^\top \Gamma u + o_p(1)$. Before discussing ΔH_1, we first define $\kappa = \frac{1}{\sqrt{m}}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\nabla h(t_{ij}, X_i;\phi^*)\{I(\in_{ij} < b_\tau^*) - \tau\}$.

Then, we have $E(\kappa) = 0$ and $Var(\kappa) = \tau(1-\tau)\Gamma$. By the Cramer-Wald Theorem and the Central Limit Theorem, we can have that $\kappa \to_d N\{0, \tau(1-\tau)\Gamma\}$.

Next, we define

$$\kappa_1 = \frac{1}{\sqrt{m}}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\nabla h(t_{ij}, X_i;\phi^*)\{I(\epsilon_{ij} < b_\tau^* + r_{ij}) - \tau\},$$

so that $\Delta H_1 = \kappa_1^\top \eta + o_p(1)$. By simple calculation, we have

$$\begin{aligned} Var(\kappa_1 - \kappa) &= \frac{1}{m} Var\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i}\nabla h(t_{ij}, X_i;\phi^*)\{I(\epsilon_{ij} < b_\tau^* + r_{ij}) - I(\epsilon_{ij} < b_\tau^*)\}\right] \\ &= \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\nabla h(t_{ij}, X_i;\phi^*)\{\nabla h(t_{ij}, X_i;\phi^*)\}^\top Var\{I(\epsilon_{ij} < b_\tau^* + r_{ij}) - I(\epsilon_{ij} < b_\tau^*)\} \\ &\le \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\nabla h(t_{ij}, X_i;\phi^*)\{\nabla h(t_{ij}, X_i;\phi^*)\}^\top E|I(\epsilon_{ij} < b_\tau^* + r_{ij}) - I(\epsilon_{ij} < b_\tau^*)| \\ &\le \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\nabla h(t_{ij}, X_i;\phi^*)\{\nabla h(t_{ij}, X_i;\phi^*)\}^\top \left\{G(b_\tau^* + r_{ij}) - G(b_\tau^*)\right\} = o_p(1). \end{aligned}$$

Thus $\kappa_1 \to_p \kappa$. By Slutsky's theorem, $\kappa_1 \to_d N\{0, \tau(1-\tau)\Gamma\}$. Then, we can have that

$$\Delta H = \frac{1}{2}g(b_\tau^*)u^\top \Gamma u + \kappa_1^\top u + o_p(1).$$

By the epi-convergence results of [28], $\hat{u} \to_d - g(b_\tau^*)^{-1}\Gamma^{-1}\kappa_1$. Finally, the asymptotic normality is proved $\hat{u} \to_d N\{0, g(b_\tau^*)^{-2}\tau(1-\tau)\Gamma^{-1}\}$.

Since $\| \hat{\eta}_k(u) - \eta_k(u) \|_2 \leq \| \hat{\eta}_k(u) - \gamma_k^\top B(u) \|_2 + \| \gamma_k^\top B(u) - \eta_k(u) \|_2$, we have

$$\begin{aligned} \| \hat{\eta}_k(u) - \gamma_k^\top B(u) \|_2 &= \{E(\hat{\gamma}_k^\top B(u) - \gamma_k^\top B(u))^2\}^{1/2} \\ &= [E\{tr[(\hat{\gamma}_k - \gamma_k)^\top B(u) B(u)^\top (\hat{\gamma}_k - \gamma_k)]\}]^{1/2} \\ &= [tr\{E(B(u)B(u)^\top) E(\hat{\gamma}_k - \gamma_k)(\hat{\gamma}_k - \gamma_k)^\top\}]^{1/2} \\ &= O_p[\{g(b_\tau^*)^{-2}\tau(1-\tau)\}^{1/2} (L/\textstyle\sum_{i=1}^n m_i)^{1/2}] = O_p\{(L/m)^{1/2}\}. \end{aligned}$$

By the Lemma 1, $\| \hat{\eta}_k(u) - \eta_k(u) \|_2 = O_p\{(L/m)^{1/2}\}, k = 1, \ldots, p+1$. Thus, we can get

$$\| \log \hat{\lambda}_0(u) - \log \lambda_0(u) \|_2 = O_p\{(L/m)^{1/2}\},$$

and

$$\| \hat{\beta}_k(u) - \beta_k(u) \|_2 = O_p\{(L/m)^{1/2}\}, k = 1, \cdots, p.$$

Author Contributions

Conceptualization: Weiwei Wang.

Formal analysis: Yijun Wang.

Funding acquisition: Weiwei Wang.

Methodology: Yijun Wang.

Software: Weiwei Wang.

Writing – original draft: Weiwei Wang.

Writing – review & editing: Weiwei Wang.

References

1. Diggle PJ, Liang KY, Zeger SL. The analysis of longitudinal data. Oxford University Press New York; 1994.
2. Sun Y. Estimation of semiparametric regression model with longitudinal data. Lifetime Data Analysis. 2010; 16(2):271–298. https://doi.org/10.1007/s10985-009-9136-2 PMID: 19890712
3. Nielsen JD, Dean CB. Clustered mixed nonhomogeneous Poisson process spline models for the analysis of recurrentevent panel data. Biometrics. 2008; 64(3):751–761. https://doi.org/10.1111/j.1541-0420.2007.00940.x PMID: 18047528
4. Lu M, Zhang Y, Huang J. Semiparametric estimation methods for panel count data using monotone B-splines. Journal of the American Statistical Association. 2009; 104(487):1060–1070. https://doi.org/10.1198/jasa.2009.tm08086
5. He X, Tong X, Sun J. Semiparametric analysis of panel count data with correlated observation and follow-up times. Lifetime Data Analysis. 2009; 15(2):177–196. https://doi.org/10.1007/s10985-008-9105-1 PMID: 19082711
6. Zhao X, Tong X. Semiparametric regression analysis of panel count data with informative observation times. Computational Statistics and Data Analysis. 2011; 55(1):291–300. https://doi.org/10.1016/j.csda.2010.04.020
7. Zhao X, Tong X, Sun J, Azen SP. Robust estimation for panel count data with informative observation times. Computational Statistics and Data Analysis. 2013; 57(1):33–40. https://doi.org/10.1016/j.csda.2012.05.015
8. Li N, Sun L, Sun J. Semiparametric transformation models for panel count data with dependent observation process. Statistics in Biosciences. 2010; 2(22):191–210. https://doi.org/10.1007/s12561-010-9029-7

9. Li N. Semiparametric transformation models for panel count data. University of Missouri–Columbia; 2011.
10. Li N, Zhao H, Sun J. Semiparametric transformation models for panel count data with correlated observation and follow-up times. Statistics in Medicine. 2013; 32(17):3039–3054. https://doi.org/10.1002/sim.5724 PMID: 23297190
11. Sun J, Zhao X. Statistical analysis of panel count data. Springer New York; 2013.
12. Chiang CT, Wang MC. Varying-coefficient model for the occurrence rate function of recurrent events. Annals of the Institute of Statistical Mathematics. 2009; 61(1):197–213. https://doi.org/10.1007/s10463-007-0129-1
13. Sun L, Zhou X, Guo S. Marginal regression models with time-varying coefficients for recurrent event data. Statistics in Medicine. 2011; 30(18):2265–2277. https://doi.org/10.1002/sim.4260 PMID: 21590791
14. He X, Feng X, Tong X, Zhao X. Semiparametric partially linear varying coefficient models with panel count data. Lifetime Data Analysis. 2016; 23(3):1–28. PMID: 27118299
15. Zhao H, Tu W, Yu Z. A nonparametric time-varying coefficient model for panel count data. Journal of Nonparametric Statistics. 2018; 30(3):640–661. https://doi.org/10.1080/10485252.2018.1458982
16. Huang JZ, Wu CO, Zhou L. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. Biometrika. 2002; 89(1):111–128. https://doi.org/10.1093/biomet/89.1.111
17. Machado JAF, Silva JMCS. Quantiles for counts. Journal of the American Statistical Association. 2005; 100(472):1226–1237. https://doi.org/10.1198/016214505000000330
18. Guo J, Tang M, Tian M, Zhu K. Variable selection in high-dimensional partially linear additive models for composite quantile regression. Computational Statistics and Data Analysis. 2013; 65(9):56–67. https://doi.org/10.1016/j.csda.2013.03.017
19. Sun J, Wei LJ. Regression analysis of panel count data with covariate-dependent observation and censoring times. Journal of the Royal Statistical Society. 2000; 62(2):293–302. https://doi.org/10.1111/1467-9868.00232
20. Fan TH, Cheng KF. Tests and variables selection on regression analysis for massive datasets. Data & Knowledge Engineering. 2007; 63(3):811–819. https://doi.org/10.1016/j.datak.2007.05.001
21. Liquet B, Saracco J. BIG-SIR: a sliced inverse regression approach for massive data. Statistics and its Interface. 2016; 9(4):509–520. https://doi.org/10.4310/SII.2016.v9.n4.a10
22. Perry PO, Wolfe PJ. Point process modelling for directed interaction networks. Journal of the Royal Statistical Society. 2013; 75(5):821–849. https://doi.org/10.1111/rssb.12013
23. Xu A, Zhou S, Tang Y. A unified model for system reliability evaluation under dynamic operating conditions. IEEE Transactions on Reliability. 2021; 70(1):65–72. https://doi.org/10.1109/TR.2019.2948173
24. Chen F, Chen S, Ma X. Crash frequency modeling using real-time environmental and traffic data and unbalanced panel data models. International Journal of Environmental Research and Public Health. 2016; 13(6):609. https://doi.org/10.3390/ijerph13060609 PMID: 27322306
25. Chen F, Chen S, Ma X. Analysis of hourly crash likelihood using unbalanced panel data mixed logit model and real-time driving environmental big data. Journal of Safety Research. 2018; 65:153–159. https://doi.org/10.1016/j.jsr.2018.02.010 PMID: 29776524
26. Dong B, Ma X, Chen F, Chen S. Investigating the differences of single-vehicle and multivehicle accident probability using mixed logit model. Journal of Advanced Transportation. 2018, UNSP 2702360. https://doi.org/10.1155/2018/2702360
27. Knight K. Limiting distributions for L_1 regression estimators under general conditions. Annals of Statistics. 1998; 26(2):755–770. https://doi.org/10.1214/aos/1028144858
28. Knight K, Fu W. Asymptotics for lasso-type estimators. Annals of Statistics. 2000; 28(5):1356–1378.