Journal of Multivariate Analysis 167 (2018) 60–78

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Robust variable selection of joint frailty model for panel count data

Weiwei Wang [a], Xianyi Wu [a], Xiaobing Zhao [b,*], Xian Zhou [c]

[a] *School of Statistics, East China Normal University, Shanghai, China*
[b] *School of Data Sciences, Zhejiang University of Finance and Economics, Hangzhou, Zhejiang Province, China*
[c] *Department of Applied Finance and Actuarial Studies, Macquarie University, NSW, Australia*

ARTICLE INFO

Article history:
Received 13 February 2018
Available online 20 April 2018

AMS subject classifications:
62H12
62J02

Keywords:
Panel count data
Joint frailty model
Variable selection
Quantile regression

ABSTRACT

Panel count data are generated from studies that concern recurrent events or event history studies in which the subjects are observed only at specific points in time. Recently, research on panel count data has drawn considerable attention. The literature on variable selection of panel count data has so far been quite limited. In this paper, a robust variable selection approach based on the quantile regression function in a joint frailty model is proposed to analyze panel count data. A three-step estimation method is introduced to estimate the coefficients and unknown functions. Consistency and oracle properties are established under some mild regularity conditions. Simulations are used to assess the proposed estimation method. Bladder tumor cancer data are also re-analyzed as an illustration.

1. Introduction

The statistical analysis of event history data has attracted a great deal of attention and various methods can be found in the literature. Of particular interest in this paper are panel count data, which are obtained at discrete observation time points and only include the numbers of recurrent events occurred at such time points. As continuous observation is often impractical, panel count data occur in various fields including medical research, insurance studies, reliability and tumorigenicity experiences. Analyzing such data can be challenging because information is scarce.

In the study of panel count data, various models and approaches have been developed. Diggle et al. [2] and Sun [21] applied methods from regression analysis of longitudinal data to panel count data. Nielsen and Dean [18] estimated the semiparametric model of panel count data by using penalized splines. Lu et al. [16] suggested using the likelihood method to estimate the multiplicative model. These models all assume independence between the recurrent event process and the observation process. This independence assumption, however, may be unreasonable in practice.

He et al. [7] introduced a joint model in which two latent frailty parameters are adopted to account for the correlations between the two processes and they developed a three-step parameter estimation approach. Zhao and Tong [32] proposed a robust joint model with an unspecified correlation link function. Zhao et al. [31] considered a joint model with observation processes and a terminal event. Li et al. [14] investigated univariate panel count data with correlated observations by semiparametric transformation models. Furthermore, Li [13] extended semiparametric transformation models with correlated observation process to the multivariate case. Li et al. [15] proposed semiparametric transformation models with correlated observation times and follow-up times. Zhao et al. [33] developed a robust estimation method of the

[*] Corresponding author.
E-mail addresses: 52153001015@ecnu.cn (W. Wang), xywu@stat.ecnu.edu.cn (X. Wu), maxbzhao@zufe.edu.cn (X. Zhao), xian.zhou@mq.edu.au (X. Zhou).

https://doi.org/10.1016/j.jmva.2018.04.003

semiparametric transformation models. More discussion regarding this type of data can be found in the book by Sun and Zhao [23].

To select significant covariates, variable selection approaches have been extensively studied by many authors in regression analysis. The least absolute shrinkage and selection operator (Lasso) was presented by Tibshirani [24]. Fan and Li [4] developed the nonconcave penalized likelihood approach (SCAD) for variable selection and showed that their method has an oracle property under some regularity conditions. Zou [35] developed an Adaptive Lasso and showed its oracle property. Zhang [29] proposed a minmax concave penalty (MCP). However, research on variable selection on panel count data is limited. Recently, Tong et al. [25] considered the estimation of panel count data by adding the SCAD penalty function to the estimating function. Zhang et al. [30] also looked at the multivariate case.

In the analysis of regression models, the presence of heavy tail distributions and outliers can cause inefficient estimation. To overcome these difficulties, robust estimators have been extensively studied by many authors, especially the quantile regression method. Quantile regression contains more information about the distribution shape and can be used to measure the effect of variables not only in the center of the distribution, but also in the upper and lower tails. It has been widely used in regression modeling. Little has been done, however, for panel count data with quantile regression model, which is considered in this paper. As panel count data are discrete, we cannot directly use the quantile regression method to get robust estimation. Instead, a smoothing technique can be applied to panel count data and then the general method of quantile regression can be used to obtain robust estimation.

In this paper, a robust variable selection method is developed for the joint model proposed by He et al. [7]. A three-step estimation method is developed for model estimation. In the first step, the observation process is estimated by a nonparametric estimation for the baseline function and penalized estimation for the covariates. In the second step, an EM algorithm and penalized estimation are used for the follow-up process. In the third step, a robust variable selection method based on the quantile regression function is developed for the recurrent event process. As panel count data are discrete, a smoothing technique is used to get the continization of the data. Then a spline basis expansion is applied to approximate the unknown baseline function. Finally, a penalized estimation based on quantile regression is proposed to estimate the parameters of interest.

In addition to developing the three-step estimation method, we further study the consistency and oracle properties of the estimators. The robust variable selection method based on quantile regression is new to the literature of panel count data. The main innovation of this paper lies in the estimation of the panel count data under different quantiles, and the variable selection method to select the significant variables simultaneously.

The structure of the paper is as follows. Some notions and the joint model are specified in Section 2. We introduce the three-step estimation method for the joint model of panel count data in Section 3. Section 4 summaries the consistency and oracle properties of the estimators. Some simulation studies are presented in Section 5. In Section 6, bladder cancer data are analyzed with the proposed approach. A summary and discussion are given in Section 7.

2. Model specification

We first introduce the basic structure of panel count data. For each subject $i \in \{1, \ldots, n\}$, let $N_i(t)$ denote the number of occurrences of the event at or before time t, and $\tilde{H}_i(t)$ represent a counting process which jumps only at the potential observation times $t_{i,1} < t_{i,2} < \cdots$ on $N_i(t)$. Furthermore, there exist a potential censoring time C_i^* for subject i and an endpoint T of observation in the recurrent event process. We can only observe $C_i = \min(C_i^*, T)$ and $\delta_i = \mathbf{1}(C_i^* \le T)$. Furthermore, C_i^* may be correlated with $N_i(t)$ and $\tilde{H}_i(t)$, but T is independent of them. Define

$$H_i(t) = \tilde{H}_i\{\min(t, C_i)\} = \sum_{j=1}^{m_i} \mathbf{1}(t_{i,j} \le t),$$

which is the actual observation process of subject $i \in \{1, \ldots, n\}$, where $m_i = \tilde{H}_i(C_i)$. Let Z be a $p \times 1$ covariates, and Z_i the value of Z for subject i. Therefore, we have the dataset $\{H_i(t), N_i(t)dH_i(t), C_i, \delta_i, Z_i : t \ge 0, i \in \{1, \ldots, n\}\}$. To build a model on $N_i(t)$, $\tilde{H}_i(t)$ and C_i^*, the approach of He et al. [7] is defined as follows.

Suppose that two independent latent variables u_i and v_i are involved in the model.

(A) Given Z_i, u_i and v_i, $N_i(t)$, assume that $\tilde{H}_i(t)$ and C_i^* are mutually independent.

(B) Given Z_i and u_i, assume that $\tilde{H}_i(t)$ follows a non-homogeneous Poisson process with intensity function

$$\lambda_{ih}(t|Z_i, u_i) = \lambda_{0h}(t) \exp(\alpha^\top Z_i + u_i), \tag{1}$$

where α is a $p \times 1$ vector of parameters (coefficients of Z) and $\lambda_{0h}(t)$ is an unknown continuous baseline function with $\Lambda_{0h}(t) = \int_0^t \lambda_{0h}(s)ds$. Assume $\Lambda_{oh}(T) = 1$ for model identification. Moreover, the random variable u_i is assumed to be symmetric about zero, but its distribution need not be specified. Further assume that $\mathrm{E}(u_i|Z_i) = \mathrm{E}(u_i)$. After integrating out u_i in Eq. (1), the intensity function in (1) is transformed into a marginal rate function, viz. $\lambda_{ih}(t|Z_i) = \lambda_{0h}(t) \exp[\alpha^\top Z_i + \ln \mathrm{E}\{\exp(u_i)\}]$.