Computational Statistics and Data Analysis 97 (2016) 47–59

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Semiparametric regression analysis of panel count data allowing for within-subject correlation

Bin Yao [a,*], Lianming Wang [a], Xin He [b]

[a] Department of Statistics, University of South Carolina, Columbia, SC 29208, USA

[b] Department of Epidemiology and Biostatistics, University of Maryland, College Park, MD 20742, USA

ARTICLE INFO

Article history:
Received 25 June 2014
Received in revised form 28 October 2015
Accepted 26 November 2015
Available online 5 December 2015

Keywords:
EM algorithm
Gamma frailty
Monotone splines
Panel count data
Poisson process

ABSTRACT

In this paper, a maximum likelihood approach is proposed for analyzing panel count data under the gamma frailty non-homogeneous Poisson process model. The approach allows one to estimate the baseline mean function and the regression parameters jointly while taking the within-subject correlation into account. The within-subject correlation is quantified explicitly by Pearson's correlation coefficient. Monotone splines are adopted to approximate the unspecified nondecreasing baseline mean function in the model. An expectation–maximization (EM) algorithm is derived to facilitate the computation by exploiting a data augmentation based on Poisson latent variables. The EM algorithm is robust to initial values, easy to implement, converges fast, and provides closed-form variance estimates. It can be also applied to the non-homogeneous Poisson model without frailty. The proposed approach is evaluated through simulations and illustrated by two real-life examples coming from a skin cancer study and a bladder tumor study. A companion R package `PCDSpline` has been developed and is available on R CRAN for public use.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Panel count data often arise in longitudinal prospective studies involving recurrent events that are only detected and recorded at periodic observation/assessment times. For each subject, observations are taken at finite discrete time points, and only the number of events that occur between consecutive observation times is known. Furthermore, the set of observation times can vary from subject to subject. Areas that usually produce panel count data include demographical studies, epidemiological studies, medical follow-up studies, oncology clinical trials, and reliability studies (Kalbfleisch and Lawless, 1985; Sun and Kalbfleisch, 1995; Wellner and Zhang, 2000; Sun and Zhao, 2013).

Many approaches have been proposed to analyze panel count data based on the counting process techniques. When no covariates are considered, inferences are focused on estimating the mean function of the counting process. For this purpose, Sun and Kalbfleisch (1995) constructed an isotonic regression estimator. Wellner and Zhang (2000) proposed two estimators by maximizing the pseudolikelihood and likelihood functions under the non-homogeneous Poisson process. Lu et al. (2007) proposed likelihood-based estimators with the mean function being approximated by the monotone splines of Ramsay (1988) and showed that their spline-based estimators are more efficient than those in Wellner and Zhang (2000). Regarding the comparison of the mean functions for different populations, Sun and Fang (2003), Zhang (2006), and Balakrishnan and Zhao (2009) proposed different nonparametric tests for univariate panel count data. Li et al. (2014) developed nonparametric tests for multivariate panel count data.

* Corresponding author.
E-mail address: yaob@email.sc.edu (B. Yao).

http://dx.doi.org/10.1016/j.csda.2015.11.017
0167-9473/© 2015 Elsevier B.V. All rights reserved.

When covariates are available, semiparametric regression analysis is widely used to examine the covariate effects on the response as well as the estimation of the mean function. Among others, Sun and Wei (2000) developed estimation procedures with time-dependent covariates on both observational and censoring processes. Zhang (2002) proposed a pseudolikelihood approach and Wellner and Zhang (2007) studied both pseudolikelihood and likelihood methods under the non-homogeneous Poisson process model. Hu et al. (2003) proposed two estimation approaches with different assumptions on the observational process. Lu et al. (2009) modeled the baseline mean function with monotone B-splines and established the asymptotic properties of their spline-based estimators. He et al. (2008) considered the regression analysis of multivariate panel count data. There are also many approaches developed for the cases that the recurrent event process and the observational process are dependent. For such work, we refer to Sun and Zhao (2013) for a comprehensive review.

In this paper, we study semiparametric regression analysis of panel count data while taking into account within-subject correlation. Such within-subject correlation naturally exists because panel counts are repeatedly measured from the same subject. Ignoring such within-subject correlation may lead to serious problems as shown in our simulation studies. Existing work allowing for within-subject correlation in panel count data is limited. Zhang and Jamshidian (2003) proposed an EM algorithm based on the gamma frailty Poisson model but without incorporating covariates. Recently, Hua and Zhang (2012) developed a spline-based semiparametric projected generalized estimating equation approach and modeled the baseline mean function with monotone cubic B-splines. Their method does not require the Poisson assumption, and their estimating equation can be regarded as the score equation of the marginal likelihood under the gamma frailty Poisson model when the frailty variance parameter is known. Hua et al. (2014) essentially adopted the same computational algorithm as Hua and Zhang (2012) under the gamma frailty Poisson model and established the asymptotic properties of their spline-based estimators. Although their models allow for handling within-subject correlation, none of the above papers derived or estimated such within-subject correlations.

In this paper, a maximum likelihood approach is proposed for analyzing panel count data under the gamma frailty Poisson model when there is within-subject correlation. The within-subject correlation is quantified by Pearson's correlation coefficient in an explicit form. Monotone splines of Ramsay (1988) is adopted to model the baseline mean function, and all the parameters are estimated jointly through an efficient EM algorithm. The EM algorithm is robust to initial values, converges fast, and provides variance estimates of all parameters in closed form. Our approach has a good performance under the gamma frailty model and can be also applied to panel count data where there is no within-subject correlation. An R package `PCDSpline` has been developed based on our method and is now available on R CRAN for public use. Discussions on the differences between our method and that in Hua et al. (2014) can be found in the later sections.

The remainder of the article is organized as follows. Section 2 presents some notations, the model, the observed likelihood, and the modeling of the baseline mean function with monotone splines. Section 3 gives details of our proposed approach including a data augmentation, the derivation of an EM algorithm, the variance estimates, and also a brief discussion of a simplified version adapted for the case where there is no within-subject correlation. Section 4 evaluates the performance of our approach through simulations. Section 5 provides two real-life applications from a skin cancer study and a bladder tumor study. Section 6 concludes with some discussions.

2. The proposed model

2.1. Notation, model, and likelihood

Consider a study that consists of n independent subjects. We assume that the observational process and the recurrent event process are conditionally independent given covariates. Let $\boldsymbol{x}_i$ denote a vector of $p \times 1$ time-independent covariates and $\{t_{ij},\ j = 1, \ldots, K_i\}$ denote the actual observational times for subject i, where K_i is the number of observations and t_{iK_i} is the last observation time. Let $N_i(t)$ denote the counting process for subject i, and this process is observed only at t_{ij}'s. In order to account for the within-subject correlation, we propose a gamma frailty non-homogeneous Poisson process model for the recurrent event process $N_i(t)$. Specifically, conditional on ϕ_i, the frailty associated with subject i, $N_i(t)$ is a non-homogeneous Poisson process with mean function $\mu_0(t)\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\phi_i$, where $\mu_0(t)$ is an unspecified nondecreasing baseline mean function with $\mu_0(0) = 0$ and ϕ_i's are independently and identically distributed from $\mathcal{G}a(\nu, \nu)$ with mean 1 and variance ν^{-1}. This model implies

$$N_i(t)|\phi_i \sim \mathcal{P}\{\mu_0(t)\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\phi_i\}$$

for any $t \geq 0$, where $\mathcal{P}(a)$ denotes the Poisson distribution with mean a. In this model the mean constraint of the frailty distribution is made to avoid non-identifiability because $\mu_0(\cdot)$ is unspecified. Under the proposed gamma frailty Poisson process model, $\mu_0(\cdot)$ is the conditional baseline mean function of the recurrent event process given the frailty but can also be interpreted as the marginal baseline mean function since

$$\mathrm{E}\{N_i(t)\} = \mathrm{E}[\mathrm{E}\{N_i(t)|\phi_i\}] = \mathrm{E}\{\mu_0(t)\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\phi_i\} = \mu_0(t)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}).$$

Under the proposed model, the common frailty among the panel counts within the same subjects induces within-subject correlation, while the panel counts for different subjects are independent. The ϕ_i's represent the heterogeneity not explained by the covariates among the subjects, and the variance parameter ν attributes to the degree of the within-subject association

Download English Version:

https://daneshyari.com/en/article/6869352

Download Persian Version:

https://daneshyari.com/article/6869352

Daneshyari.com