Computational Statistics and Data Analysis 73 (2014) 103–111

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Nonparametric tests for panel count data with unequal observation processes

Yang Li [a], Hui Zhao [b,*], Jianguo Sun [c], KyungMann Kim [d]

[a] Department of Mathematics and Statistics, University of North Carolina-Charlotte, Charlotte, NC, 28223, USA
[b] Department of Statistics, Central China Normal University, Wuhan, 430079, PR China
[c] Department of Statistics, University of Missouri, Columbia, MO, 65211, USA
[d] Departments of Biostatistics & Medical Informatics and Statistics, University of Wisconsin, Madison, WI, 53792, USA

ARTICLE INFO

Article history:
Received 23 April 2013
Received in revised form 15 November 2013
Accepted 21 November 2013
Available online 4 December 2013

Keywords:
Nonparametric comparison
Unequal observation processes
Univariate and multivariate panel count data
Sparsely distributed data

ABSTRACT

Nonparametric comparison for panel count data is discussed. For the situation, most available approaches require that all subjects have the same observation process. However, such an assumption may not hold in reality. To address this, a new class of test procedures are proposed that allow unequal observation processes for the subjects from different treatment groups. The method applies to both univariate and multivariate panel count data. In addition, the asymptotic normality of the proposed test statistics is established and a simulation study is conducted to evaluate the finite sample properties of the proposed approach. The simulation results show that the proposed procedures work well for practical situations and in particular for sparsely distributed data. They are applied to a set of panel count data arising from a skin cancer study.

1. Introduction

Event history data concerning recurrent events are commonly encountered in medical and reliability studies. Panel count data are one class of such data that arise when study subjects can be observed only discretely (Kalbfleisch and Lawless, 1985; Sun and Zhao, 2013; Zhao et al., 2011). In this case, instead of observing the exact event times, one knows only the numbers of the events that happen between observation times. Furthermore, the observation times may vary from subject to subject and are unbalanced.

One example of panel count data is given by the skin cancer chemoprevention trial conducted by the University of Wisconsin Comprehensive Cancer Center in Madison, Wisconsin (Li et al., 2011). It is a five-year double-blinded, placebo-controlled, randomized Phase III clinical trial. One primary objective was to evaluate the effectiveness of 0.5 g/m^2/day PO difluoromethylornithine (DFMO) in reducing new skin cancers in a population of patients with a history of non-melanoma skin cancers: basal cell carcinoma and squamous cell carcinoma. The patients were scheduled to be observed every six months. As expected, however, the actual observation times differ from patient to patient and so as the follow-up times. Especially, the observation times are very sparsely distributed.

In many medical studies that produce panel count data, including the example given above, treatment comparison is one of the most asked questions. The majority of the existing test procedures assume identical observation processes across different treatment groups or involve the mean function estimation. For example, Thall and Lachin (1988) suggested transforming the problem to a multivariate comparison one by grouping panel count data to multivariate data. Sun

* Corresponding author. Tel.: +86 27 67862070; fax: +86 27 67867450.
E-mail address: hzhao@mail.ccnu.edu.cn (H. Zhao).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.11.014

and Kalbfleisch (1993), Sun and Fang (2003) and Park et al. (2007) developed model-free approaches employing the isotonic regression estimator (IRE) of the underlying mean function. Zhang (2006) and Balakrishnan and Zhao (2011) used nonparametric maximum pseudo-likelihood estimator (NPMPLE) for multi-sample comparisons. Also Balakrishnan and Zhao (2009, 2010) developed some procedures by employing the nonparametric maximum likelihood estimator (NPMLE). All of the approaches above require an identical observation process across all study subjects, which may not be feasible in practice. For this, Zhao and Sun (2011) proposed a test procedure that allows for unequal observation processes. However, their test statistics also involve the mean function estimation and they employed the IRE.

Although the mean function estimators IRE, NPMPLE or NPMLE perform well in general, we noticed that they may be biased when the data or observation times are sparsely distributed like the skin cancer data described above. In this article, we propose a new class of nonparametric test procedures that allow different observation processes without employing the mean function estimation. The new test procedure is motivated by those used for recurrent event data. Unlike the most of the test procedures listed above, the test statistics are constructed as contrasts of the sample means of the integrated weighted responses from the underlying recurrent event processes. It will be seen that the proposed test procedure performs well and in particular, for sparsely distributed data.

The remainder of the article is organized as follows. Section 2 first considers the comparison problem for univariate panel count data and presents a class of test procedures. Section 3 then generalizes the test procedure to multivariate panel count data. For both cases, the asymptotic normality of the test statistics is established. Section 4 investigates the finite sample properties of the proposed test procedures through a simulation study and Section 5 applies the methodology to the skin cancer study described above. Some concluding remarks are provided in Section 6.

2. Nonparametric tests for univariate panel count data

Consider m groups of independent subjects in a recurrent event study with total sample size n. Suppose that only panel count data are available and observation processes are different for the subjects from different groups. Specifically, assume that there are n_l subjects in the lth group, $l = 1, \ldots, m$, and let S_l denote the set of indices for the subjects in group l, where $\sum_{l=1}^{m} n_l = n$. Suppose that Z_i is a group-indicating vector associated with subject i ($i = 1, \ldots, n$). In practice, study subjects may be grouped either by treatments or some covariates of interest and each Z_i may include covariates besides group indicators. Let $Y_i(t)$ be the counting process representing the total number of recurrent event occurrences up to time t from subject i with $\mu_l(t) = E\{Y_i(t)|Z_i\}$ for $i \in S_l$. In addition, let C_i denote the censoring or follow-up time of subject i. It censors the observation times $T_{i,1} < T_{i,2} < \cdots$ in the sense that the event process $Y_i(\cdot)$ is observed only at jumps of $N_i(t) = N_i^*(C_i \wedge t)$, where $N_i^*(t) = \sum_{j=1}^{\infty}(T_{i,j} \le t)$ and $a \wedge b$ denotes the minimum of a and b. Let m_i represent the total number of observation times for subject i and τ be the longest follow-up time. To account for the fact that subjects with different covariates may have different observation processes, we assume that $N_i^*(t)$ depends on Z_i through the rate model

$$E\{dN_i^*(t)|Z_i\} = \exp(\gamma' Z_i)\lambda_0(t)dt, \tag{2.1}$$

where $\lambda_0(\cdot)$ is an unspecified continuous function and γ is a vector of unknown regression parameters. Model (2.1) implies that Z_i has a multiplicative effect on the number of observations, and $\gamma = 0$ means that the observation processes are the same. Similar proportional models have been considered by many authors including Lin et al. (2000), Sun and Wei (2000), Lin et al. (2001), Sun et al. (2005) and Li et al. (2010) among others. The adequacy of model (2.1) is relatively easy to check since the observation process provides complete data.

Unlike the observation process, the recurrent event process associated with panel count data is not continuously observed and thus its model adequacy is generally difficult to check. In this article, we focus on a treatment comparison procedure which is model-free of the recurrent event process with panel count data while model (2.1) holds. Suppose that C_i is independent of $\{Z_i, Y_i(t), N_i^*(t)\}$ and the observation process is noninformative, that is, $Y_i(t)$ and $N_i^*(t)$ are independent given Z_i. The observed data consist of $\{N_i(t), Z_i, C_i, Y_i(T_{i,1}), \ldots, Y_i(T_{i,m_i});\ 0 \le t, T_{i,m_i} \le C_i,\ i = 1, \ldots, n\}$.

Our aim is to test the hypothesis

$$H_0 : \mu_1(t) = \cdots = \mu_m(t),$$

that is, the occurrence rate of the recurrent event of interest is the same for different groups. Under model (2.1) and conditional on Z_i,

$$E\left\{\sum_{j=1}^{m_i} Y_i(T_{i,j})|Z_i\right\} = E\left\{\int_0^\tau Y_i(t)dN_i(t)|Z_i\right\} = \int_0^\tau \mu(t)G(t)\exp(\gamma' Z_i)\lambda_0(t)dt,$$

where $G(t) = P(C_i \ge t)$ and $\mu(t)$ denotes the common mean function of $Y_i(t)$ under H_0. Hence

$$E\left\{\int_0^\tau \frac{Y_i(t)dN_i(t)}{\exp(\gamma' Z_i)}|Z_i\right\} = \int_0^\tau \mu(t)G(t)\lambda_0(t)dt. \tag{2.2}$$