Journal of Multivariate Analysis 140 (2015) 209–219

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Semiparametric regression of multivariate panel count data with informative observation times

Yang Li [a,*], Xin He [b], Haiying Wang [c], Bin Zhang [d], Jianguo Sun [e]

[a] *Department of Mathematics and Statistics, University of North Carolina-Charlotte, Charlotte, NC, 28223, USA*
[b] *Department of Epidemiology and Biostatistics, University of Maryland, College Park, MD 20742, USA*
[c] *Department of Mathematics and Statistics, University of New Hampshire, Durham, NH 03824, USA*
[d] *Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA*
[e] *Department of Statistics, University of Missouri, Columbia, MO, 65211, USA*

ARTICLE INFO

Article history:
Received 13 September 2014
Available online 27 May 2015

AMS subject classifications:
62H12
62J99
62N05

Keywords:
Estimating equation
Informative censoring
Informative observation process
Multivariate panel count data

ABSTRACT

Multivariate panel count data occur in many fields such as medical and social science studies in which several outcomes of interest are measured simultaneously and repeatedly over time. When the observation times are not pre-specified, it is very likely that either the observation or follow-up times are informative about the response process. In such situations, most existing approaches either specify a dependence structure with some fixed distributions or assume independence given some covariates, which may not be true and result in misleading conclusions. In this paper, we present a joint modeling approach that allows the possible mutual correlations to be characterized by time-dependent random effects. Estimating equations are developed for the parameter estimation and the resulted estimators are shown to be consistent and asymptotically normal. The finite sample performance of the proposed estimators is assessed through a simulation study and an illustrative example from a maternal influenza immunization study on infant growth is provided.

1. Introduction

Multivariate panel count data occur in many fields such as medical and social science studies in which several related recurrent events are of interest but the responses are recorded only at discrete times. In such situations, only the numbers of events that have occurred between the observation times are known, but the exact event occurrence times are not observable. The observed data consist of two parts, one being a sequence of discrete observation times which can be regarded as realizations from an observation process and the other being the sequences of counts that the events have occurred between discrete observation times.

One example of multivariate panel count data that motivated this research is the Mother's Gift study, a prospective, controlled, double blinded, randomized trial to assess the safety and immunogenicity in pregnant women of pneumococcal vaccines, as well as the clinical effectiveness of influenza vaccine in Bangladesh [19]. Research studies have shown that respiratory illness has significant negative impact on children's weight and height gains in Bangladesh [18]. Therefore, reducing children's respiratory illness rates has become clinically important in enhancing children's growth in countries

* Corresponding author.
E-mail address: Y.Li@uncc.edu (Y. Li).

http://dx.doi.org/10.1016/j.jmva.2015.05.014
0047-259X/

without sufficient food and health resources. One objective of the study is to evaluate the effectiveness of infant Pneumococcal conjugate vaccine (PCV7) in reducing the recurrences of febrile respiratory illness, pneumonia and difficulty with breathing. In addition to usual routine vaccines, all infants received either PCV7 or Hib conjugate vaccine at 6, 10 and 14 weeks of age. The experiences of febrile respiratory illness, pneumonia and difficulty with breathing of the infants were scheduled to be recorded weekly from 6 to 24 weeks of age, but as expected, the actual observation times varied among the infants. In particular, many infants were examined more frequently when the respiratory illness occurred and skipped a couple of clinical visits when they were healthy. Therefore, the actual observation times could be informative about the recurrence processes of respiratory illness and multivariate panel count data were generated as described above. In the original analysis of the study, the chi-square test and Fisher's exact test were used to compare the laboratory-confirmed influenza occurrence rates, defined as the total number of infant influenza cases divided by the total number of infants from each treatment group. It is easy to see that this is clearly not efficient.

For the analysis of multivariate panel count data, it is apparent that one may separately carry out univariate analysis for each type of outcomes by applying an existing procedure [7,15,24,17,25]. However, such practice ignores the mutual correlation between the related outcomes and would be less efficient than a joint or multivariate analysis. To overcome this, several authors have proposed multivariate regression models. For example, Chen et al. [1] and He et al. [5] considered parametric and semiparametric methods, respectively, by assuming that the response process and the observation process are independent of each other. Li et al. [9] and Zhao et al. [23] proposed some marginal model-based approaches that allow the dependent observation processes. However, their models imply that the subjects with the same observation schedule are expected to have the same response rates for all types of recurrent events, which clearly may not be realistic in some applications. For instance, despite of observation times in the past, the proceeding observation times and longitudinal responses may both depend on individuals' current stages of disease progression. Also such correlation may vary over time and relate to the follow-up times.

In the following, we discuss regression analysis of multivariate panel count data when the response process, the observation process and the follow-up time may be mutually correlated. An easy-to-implement estimation approach is proposed and the asymptotic properties of the resulted estimators are established. The approach allows for time-dependent, arbitrary correlations. Before presenting the estimation approach, we will first introduce the notation and present the model in Section 2. Section 3 presents the estimation procedure and establishes the asymptotic properties of the resulted estimators, and Section 4 gives a procedure for model diagnostics. In Section 5, a simulation study is conducted to evaluate the finite-sample performance of the proposed estimators, and Section 6 applies the method to the data from the maternal influenza immunization study described above. The paper concludes in Section 7 with some discussion and remarks.

2. Notation and models

Consider a recurrent event study that involves p types of events. For subject i, let $Y_{ik}(t)$ denote the total number of type-k events that have occurred up to time t, $i = 1, \ldots, n$, $k = 1, \ldots, p$. Suppose that $\mathbf{Y}_i(t) = \left(Y_{i1}(t), Y_{i2}(t), \ldots, Y_{ip}(t)\right)'$ is observed only at discrete time points $\{T_{i,1}, \ldots, T_{i,m_i}\}$, where m_i represents the total number of observations on subject i. Let $N_i(t)$ represent the observation process, which gives the cumulative numbers of observation times up to time t. In practice, there usually exists a censoring or follow-up time C_i and one observes $\widetilde{N}_i(t) = N_i(t \wedge C_i)$, where $a \wedge b = \min(a, b)$. Let $\mathbf{Z_i}(t)$ denote a d-dimensional vector of covariates which is assumed to be continuously traceable in the study and denote $\mathcal{Z}_{it} = \{\mathbf{Z_i}(s), s \le t\}$.

In the following, we assume that there exists an unobserved random vector $\mathbf{b}_i(t) = \big(b_{i1}(t), \ldots, b_{ip}(t), b_{i,p+1}(t), b_{i,p+2}(t)\big)'$ that will be used to model the correlation between $\mathbf{Y}_i(t)$, $N_i(t)$ and C_i. Define $\mathcal{B}_{it} = \{\mathbf{b}_i(s), s \le t\}$ and assume that the $\mathbf{b}_i(t)$'s are independent and identically distributed, $\mathcal{B}_{it}$ is independent of $\mathcal{Z}_{it}$, and given $\mathcal{Z}_{it}$ and $\mathcal{B}_{it}$, $\mathbf{Y}_i(t)$, $N_i(t)$ and C_i are mutually independent. Also we will assume that the mean function of $Y_{ik}(t)$ has the form

$$E\{Y_{ik}(t)|\mathbf{Z_i}(t), \mathbf{b}_i(t)\} = \Lambda_{0k}(t) \exp\{\beta'\mathbf{Z_i}(t) + b_{ik}(t)\}, \quad k = 1, \ldots, p, \tag{1}$$

where β denotes a vector of d-dimensional regression coefficients and $\Lambda_{0k}(t)$ is an unknown baseline mean function. Note that for the simplicity of presentation, we have assumed that regression parameters β are the same for different types of event, and it is straightforward to generalize the methodology proposed below to the situation where covariate effects may be different on different types of recurrent events.

For the observation process $N_i(t)$, it will be assumed that it follows the following marginal rate model

$$E\{dN_i(t)|\mathbf{Z_i}(t), \mathbf{b}_i(t)\} = \exp\{\gamma'\mathbf{Z_i}(t) + b_{i,p+1}(t)\} d\mu_0(t), \tag{2}$$

where γ is a vector of unknown regression parameters and $d\mu_0(t)$ is an unknown baseline rate function. For the censoring time C_i, we will suppose that its hazard function is given by

$$\lambda_i(t|\mathbf{Z_i}(t), \mathbf{b}_i(t)) = \lambda_0(t) + \xi'\mathbf{Z_i}(t) + b_{i,p+2}(t). \tag{3}$$

Here ξ denotes the effect of covariates on the hazard function of $C_i's$ and $\lambda_0(t)$ is an unknown baseline hazard function. That is, the C_i's follow the additive hazards models [10,8,21].