Computational Statistics and Data Analysis 107 (2017) 120–130

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Semiparametric regression analysis of multivariate longitudinal data with informative observation times

Shirong Deng [a], Kin-yat Liu [b], Xingqiu Zhao [b,c,*]

[a] School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei, China

[b] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong

[c] The Hong Kong Polytechnic University, Shenzhen Research Institute, Shenzhen, China

ARTICLE INFO

Article history:
Received 22 July 2015
Received in revised form 25 March 2016
Accepted 10 October 2016
Available online 27 October 2016

Keywords:
Estimating equation
Informative observation times
Latent variable
Model checking
Multivariate longitudinal data
Semiparametric regression

ABSTRACT

Multivariate longitudinal data arises when subjects under study may experience several possible related response outcomes. This article proposed a new class of flexible semiparametric models for multivariate longitudinal data with informative observation times through latent variables and completely unspecified link functions, which allows for any functional forms of covariate effects on the intensity functions for the observation processes. A novel estimating equation approach that does not rely on forms of link functions and distributions of frailties is developed. The asymptotic properties for the resulting estimators and the model checking technique for the overall fit of the proposed models are established. The simulation results show that the proposed approach works well. The analysis of skin cancer chemoprevention trial data is provided for illustration.

1. Introduction

In many longitudinal studies, multivariate longitudinal data arise when subjects under study may experience several related events repeatedly at distinct time points during a relatively long follow-up period. An example of multivariate longitudinal data that motivated this work is a skin cancer chemoprevention trial conducted by the University of Wisconsin Comprehensive Cancer Center in Madison, Wisconsin (Lee, 2008; Li, 2011). It was a 5-year randomized, double-blinded, and placebo-controlled Phase III clinical trial. The primary objective of this trial was to evaluate the effectiveness of 0.5 g/m^2/day PO difluoromethylornithine (DFMO) in preventing new skin cancers in a population of individuals with a history of non-melanoma skin cancers: basal cell carcinoma or squamous cell carcinoma. The subjects missed scheduled visits or visited clinic on unscheduled dates. At each visit, the number of occurrences of both basal cell carcinoma and squamous cell carcinoma since the previous visit were recorded.

In the irregularly observed longitudinal data analysis, there are two important processes involved: the response process and the observation process. A basic assumption behind the usual methods is that the observation times are independent of response variable, completely or given covariates, i.e., the observation process is noninformative (e.g., Lin and Ying, 2001; Welsh et al., 2002). However, this assumption may be violated in many applications. Such as the skin cancer study, Li et al. (2011) and Zhang et al. (2013) have verified that the clinical visit times contain some relevant information about the recurrence processes of two cancers. We call these response-dependent visit times as informative observation times.

* Corresponding author at: Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong.
E-mail address: xingqiu.zhao@polyu.edu.hk (X. Zhao).

http://dx.doi.org/10.1016/j.csda.2016.10.006
0167-9473/

Thus it is very necessary to incorporate the relationship between the response process and the observation process into longitudinal data models.

For the univariate longitudinal data analysis with informative observation times, two methods have been developed. One is the conditional modeling approach (Sun et al., 2005), which obviously characterized the dependence of the response process and the observation times. Another one is the frailty-based approach proposed by Sun et al. (2007), Liang et al. (2009), Zhao et al. (2012), and Zhou et al. (2013) among others. For example, Sun et al. (2007) used a shared latent variable or frailty to characterize the correlation between the response process and the observation times with informative censoring times. Liang et al. (2009) modeled the longitudinal data with informative observation times via two latent variables that satisfied a linear relationship where the distributional assumption for a latent variable is required. Zhao et al. (2012) considered more general joint models using a completely unspecified link function and a latent variable to characterize the correlations between the response process and the observation process, and developed estimating equation approaches. Zhou et al. (2013) considered a semiparametric mixed random effect model for the response process in the presence of informative observation and censoring times.

For the multivariate longitudinal data with informative observation times, one additional issue involved is the correlation among different types of the response processes. For the analysis of such complex data, the existing research mainly focuses on its special case where each longitudinal response process is a counting process. For example, Li et al. (2011), Zhang et al. (2013), Zhao et al. (2013) and Li et al. (2015) proposed different semiparametric regression models that allow the recurrent event process and the observation process to be correlated, by leaving the dependence structures for related types of panel count processes completely unspecified.

In this paper, motivated by Zhao et al. (2012), we propose a semiparametric marginal modeling approach for the multivariate longitudinal data with informative observation times through latent variables and different completely unspecified link functions to characterize different correlations between each type of response process and the corresponding observation process. An important advantage for the modeling is the nonrestrictive condition on the correlation between different types of response process and different correlations between each type of response process and the corresponding observation process.

The remainder of this paper is organized as follows. We begin in Section 2 by introducing notation and describing models for multivariate longitudinal data with informative observation times. In Section 3 an estimating equation approach is developed to estimate the regression parameters involved in the proposed models, and the asymptotic properties for the resulting estimators are given in this section. In Section 4, we discuss the model checking technique for goodness of fit of our models. The simulation results are presented in Section 5 to assess the finite-sample performance of the proposed inference procedure, and the analysis of skin cancer chemoprevention trial data is provided to illustrate the proposed method in Section 6. Some concluding remarks are made in Section 7.

2. Statistical model

Consider a longitudinal study that consists of n independent subjects and suppose that each subject may experience K different types of longitudinal outcomes. For subject i, let $Y_{ik}(t)$ denote the longitudinal response process with type k and suppose that $Y_{ik}(t)$ is observed at distinct time points $0 < T_{ik,1} < T_{ik,2} < \cdots < T_{ik,m_{ik}}$, where m_{ik} is the potential or scheduled number of observations on subject i with respect to the kth longitudinal response variable. Let $\mathbf{X}_i$ be a p-dimensional vector of covariates and C_i the follow-up or censoring time for subject i, $i = 1, \ldots, n$. Note that here for the simplicity of presentation, we assume that $\mathbf{X}_i$ and C_i are the same for different types of longitudinal response. The inference approach proposed below can be easily extended to the situation where there exists different covariates and follow-up or censoring times for different responses. Define $N_{ik}(t) = \sum_{j=1}^{m_{ik}} I\left(T_{ik,j} \le t\right)$, where $I(\cdot)$ is the indicator function. Then $\tilde{N}_{ik}(t) = N_{ik}(t \wedge C_i)$ is a counting process characterizing the number of observation times on subject i with respect to the kth longitudinal response variable up to time t. Then the process $Y_{ik}(t)$ is observed only at the time points where $\tilde{N}_{ik}(t)$ jumps.

For the analysis, suppose that $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iK})'$ is an unobserved random vector independent of $\mathbf{X}_i$ with Z_{ik} being positive, and assume that given $\mathbf{X}_i$ and $\mathbf{Z}_i$, $Y_{ik}(t)$ follows the marginal model

$$E\{Y_{ik}(t)|\mathbf{X}_i, \mathbf{Z}_i\} = \mu_{0k}(t) + \beta'\mathbf{X}_i + h_k(Z_{ik}), \tag{1}$$

where $\mu_{0k}(t)$ is an unknown baseline mean function, β is a p-dimensional vector of unknown regression parameters, and $h_k(\cdot)$ is a completely unspecified function with $E\{h_k(Z_{ik})\} = 0$ for identifiability. The condition $E\{h_k(Z_{ik})\} = 0$ yields that $E\{Y_{ik}(t)|\mathbf{X}_i\} = \mu_{0k}(t) + \beta'\mathbf{X}_i$ such that the uniqueness of $\mu_{0k}(t)$ and β can be ensured. Model (1) assumes that the baseline mean functions can be different for different types of longitudinal responses, however, the effects of covariates on different types of longitudinal responses are the same for the simplicity of presentation. The correlations among the K longitudinal response processes are characterized by a K-dimensional vector of unobserved frailties, where distributions of frailties are free. So, the correlation structure of longitudinal response processes is unspecified. The goal here is to estimate regression parameter β.

Give $\mathbf{X}_i$ and $\mathbf{Z}_i$, we assume that $N_{ik}(t)$ satisfies the following rate model

$$E\{dN_{ik}(t) \mid \mathbf{X}_i, \mathbf{Z}_i\} = Z_{ik}g_k(\mathbf{X}_i)d\Lambda_{0k}(t), \tag{2}$$