Journal of Multivariate Analysis 142 (2015) 144–166

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Partially linear transformation models with varying coefficients for multivariate failure time data

Zhiping Qiu [a,b,*], Yong Zhou [b,c]

[a] School of Mathematical Sciences, Huaqiao University, Quanzhou, 362021, China

[b] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, 200433, China

[c] Academy of Mathematics and Systems Science, Chinese Academy of Science, Beijing, 100190, China

ARTICLE INFO

Article history:
Received 11 October 2013
Available online 3 September 2015

AMS subject classifications:
62H12

Keywords:
Transformation model
Multivariate failure time
Estimating equations
Local polynomials
Varying-coefficients
Delete-a-group jackknife

ABSTRACT

This paper studies the estimation and inference of varying coefficients and parameters in the partially linear transformation models for multivariate failure time data. A profile martingale-based estimating method that includes global and local estimating equations is proposed. Asymptotic properties of the estimators are established. Some numerical simulations are given to show the performance of the estimation method in finite-sample situation. In order to reduce the computational burden, a simple and useful one-step estimator method is used. We further suggest a delete-a-group jackknife method to estimate asymptotic variance of estimators. A real data set from the Busselton Population Health Surveys is analyzed to illustrate the proposed methods.

1. Introduction

The proportional hazards model [14] has been fully explored in theory and extensively used in practice. However, the proportional hazards assumption is too restrictive and may not be true in some medical studies. Alternative useful model is linear transformation models which is commonly used in survival analysis and has recently attracted considerable attention [13]. Let T be the survival time and $\boldsymbol{Z}$ be a $p \times 1$ vector of covariate. The linear transformation models assume that

$$H(T) = -\boldsymbol{\beta}'\boldsymbol{Z} + \varepsilon,$$

where $H(\cdot)$ is a completely unspecified strictly increasing function with $H(0) = -\infty$, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients, and the error term ε has a known continuous distribution that is independent of $\boldsymbol{Z}$. There are many results about the statistical analysis of the linear transformation models. Cheng, Wei and Ying [11,12] proposed the inverse censoring probability weighted (ICPW) method to estimate the unknown parameter $\boldsymbol{\beta}$, and predicted the survival probabilities of the failure time; Chen, Jin and Ying [9] proposed a martingale-based estimating equations method to simultaneously estimate H and $\boldsymbol{\beta}$; More recently, Zeng and Lin [31] studied a more general form of the linear transformation models, by use of maximum likelihood method, that can handle the time-dependent covariate case, among others.

For the linear transformation models, the basic assumption is the effect of covariates is linear, but it may be inadequate in capturing the relationship between the covariates and survival time in practice. For example, in the experimental study

* Corresponding author at: School of Mathematical Sciences, Huaqiao University, Quanzhou, 362021, China.
E-mail addresses: zpqiu128@163.com, qzp@hqu.edu.cn (Z. Qiu).

http://dx.doi.org/10.1016/j.jmva.2015.08.008
0047-259X/

of lung cancer [19], statistician found out that the effect of the covariate age on the survival time is U shape. In the woman's health study of the New York University, it was found that the effect of sex hormones on growth of breast cancer also has very strong nonlinearity. Recently, Lu and Zhang [23] studied the partially linear transformation models. They proposed a martingale-based method which includes global and local estimating equations for the estimation of parametric and nonparametric covariate effects.

In the general statistical analysis and inference, one commonly assumes, for convenience, the effect of the covariates to be constant, ignoring their dynamic feature and often unable to fit the data very well. However, a varying-coefficient model, serving as an important semiparametric model, has played an increasingly important role in statistical analysis and has became more and more popular in data analysis. This model not only can deal with the interaction between variables but also can describe the dynamic effect of some covariates. Therefore, more recently, Chen and Tong [10] studied further the varying-coefficient transformation models according to the method of spline smoothing.

Multivariate failure time data are frequently observed in the fields of biology, biomedical studies, society and economics, etc. Some examples are: the survival study of married couple living in the common environment; the inference about survival in a sample of siblings or litter mates who share a common genetic makeup; the repeated breakdowns of a certain type of machinery in industrial reliability, and so on. The main feature of the multivariate failure time data is that the survival times within the same cluster might be correlated. Therefore, how to overcome the correlation and make an efficient statistical analysis and inference for this type of data is a great challenge in semiparametric models.

There are two methods for the statistical analysis of the multivariate failure time data: frailty models and marginal hazard models. The frailty models consider the conditional hazard given the unobservable frailty random variable. It is very useful when we are interested in the association of the failure types within a subject, refer to Clayton and Cuzick [13]; Vaupel, Manton and Stallard [29]; Anderson and Louis [1] and Fan and Li [16]; et al. However, when the correlation between the observations within the same cluster is not of interest, the marginal hazard model approach is more appealing. It does not specify the correlation between the observations and uses an "independent working model" assumption to deal with the data, e.g., [30,22,2,5,6,4,3,26]. In this paper we used the second method to establish the model to fit the multivariate failure time data. The literature related to the statistical analysis of marginal model with multivariate failure time data studied mainly the Cox models [2,4,3]. To the best of our knowledge, there are limited papers that studied the transformation models with multivariate failure time data.

New technical challenge arises in dealing with within-cluster dependence and the varying effects of an exposure variable for partially linear transformation models with varying coefficients. We propose combining global and local estimating equations to naturally handle all these difficulties. The global and local estimating equations in our setting are more sophisticated than that based on the time-varying model. The seminal work of the local estimating equation approach about the nonparametric regression problems is proposed by Carroll, Ruppert and Welsh [8]. The other challenge for partially linear transformation models with varying coefficients is to estimate several nonparametric functions simultaneously and results in complicated computing problems.

In this paper, we consider the partially linear transformation models with varying-coefficients for multivariate failure time data. In Section 2 we propose a martingale-based estimation method which includes the global and local estimating equations to estimate the parametric and nonparametric part respectively. In Section 3 we discuss the asymptotic properties of the proposed estimator. In Section 4 we propose a delete-a-group jackknife method to estimate the asymptotic variance of estimators. In Section 5 we examine the performance of the proposed estimator, in finite sample, by numerical simulations. Further a real data from the Busselton Population Health Surveys is analyzed to illustrate the proposed methods. Finally, we give concluding remarks in Section 6 and the related proofs of the theorems are derived in the Appendix.

2. Model and estimation methods

Without loss of generality and for simplicity, we suppose that there are n clusters from an underlying population and each cluster has K members. But the non-equal members for each cluster can be handled in the same way. For $i = 1, \ldots, n, j = 1, \ldots, K$, let T_{ij} denote the failure time of the jth member in the ith cluster, C_{ij} be the corresponding censoring time, and $\tilde{T}_{ij} = \min(T_{ij}, C_{ij})$ be the observed survival time. Let $\delta_{ij} = I(T_{ij} \le C_{ij})$ be an indicator which is equal to 1 if $\tilde{T}_{ij}$ is a failure time and 0 otherwise. Furthermore, the censoring time C_{ij} is assumed to be independent of the failure time T_{ij} conditional on the covariates $\boldsymbol{Z}_{ij}$, $\boldsymbol{V}_{ij}$ and W_{ij} (that is the so-called "independent censoring scheme"). Denote the observed data as $\{(\tilde{T}_{ij}, \delta_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{V}_{ij}, W_{ij}) : i = 1, 2, \ldots, n, j = 1, 2, \ldots, K\}$, which is an i.i.d. sample from the population $\{(\tilde{T}_j, \delta_j, \boldsymbol{Z}_j, \boldsymbol{V}_j, W_j) : j = 1, 2, \ldots, K\}$ with $\tilde{T}_j = \min(T_j, C_j)$ and $\delta_j = I(T_j \le C_j)$.

We consider the following partially linear transformation models with varying coefficients

$$H(T_{ij}) = -\boldsymbol{\beta}^T \boldsymbol{Z}_{ij} - \boldsymbol{\phi}^T(W_{ij})\boldsymbol{V}_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, n,\ j = 1, \ldots, K, \tag{2.1}$$

where $\boldsymbol{Z}_{ij} = (Z_{ij1}, Z_{ij2}, \ldots, Z_{ijp})^T$ is $p \times 1$ vector of time-independent covariates that has linear effect on the survival time, $\boldsymbol{V}_{ij} = (V_{ij1}, \ldots, V_{ijq})^T$ is $q \times 1$ vector of covariates that may interact with some exposure covariate W_{ij}. $H(\cdot)$ is an unspecified strictly increasing function with $H(0) = -\infty$, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients, $\boldsymbol{\phi}(\cdot)$ is a vector of unspecified coefficient functions with $\boldsymbol{\phi}(0) = 0$, and ε_{ij} is the error term with a known continuous distribution that