Journal of Multivariate Analysis 147 (2016) 285–294

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Local linear regression on correlated survival data

Zhezhen Jin [a,*], Wenqing He [b]

[a] Department of Biostatistics, Columbia University, New York, NY 10032, USA

[b] Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada N6A 5B7

ARTICLE INFO

Article history:
Received 2 January 2015
Available online 15 February 2016

AMS subject classifications:
62G08
62N01
62H99

Keywords:
Asymptotic bias
Correlated survival data
Kernel function
Local linear regression
Mean squared error
Nonparametric curve estimation
Unbiased data transformation

ABSTRACT

Correlated survival data arise in many contexts, and the regression analysis of such data is often of interest in practice. In this paper, we study a weighted local linear regression method for the analysis of correlated censored data, which is a natural extension of classical nonparametric regression that models directly the effect of covariates on survival time, using an unknown smooth nonparametric function. The estimation and inference are based on local linear regression and a class of unbiased data transformations. The most important feature of the proposed method is to weight local observations with local variance, which is the key to improve the estimation efficiency. We derive the asymptotic properties of the resulting estimator and show that the asymptotic variance of the nonparametric estimator is minimized with the correct specification of correlation structure. We evaluate the performance of the proposed method using simulation studies, and illustrate the proposed method with an analysis of data from the Busselton Health Study.

1. Introduction

Correlated survival data arise from many contexts, such as in studies on survival or time to occurrence of a disease with familial data, and in clinical trials with the occurrence of multiple events. Analysis of such data is challenging due to censoring and correlation among the survival times. The commonly used methods are based on models with the concept of intensity or hazard functions, namely: intensity process models, e.g., Anderson and Gill [1], Prentice, Williams and Peterson [22]; frailty models, e.g., Hougaard [10]; and marginal proportional hazards models, e.g., Wei, Lin and Weissfeld [27], Cai and Prentice [3]. Alternative approaches, such as the marginal accelerated failure time models, have also been studied, e.g., Lin and Wei [20], Lee, Wei and Ying [17], Lin, Wei and Ying [21] and Jin, Lin and Ying [11,12]. The essence of these available approaches is to model the covariate effects parametrically. However, the validity of parametric modeling of covariate effects is often difficult to check in practice. A nonparametric regression approach offers a flexible way of modeling the covariate effects.

Most nonparametric regression models have been developed for univariate censored data under the random censoring assumption. In particular, Tibshirani and Hastie [25] studied nonparametric regression estimator with kernel smoothing on local partial likelihood, Hastie and Tibshirani [8] and Gray [7] developed nonparametric regression estimators with smoothing spline methods in proportional hazards models, and Fan and Gijbels [6], Singh and Lu [24] studied nonparametric regression estimators based on the local polynomial approach, using a class of unbiased data transformations.

There is extensive literature on nonparametric regression for the analysis of correlated data without censoring; see Severini and Staniswalis [23], Lin and Carroll [19], Wang [26], Chen and Jin [5], and many others. In contrast, there is little literature on nonparametric regression for the analysis of correlated survival data. One attempt was made in Yu and

* Corresponding author.
E-mail addresses: zj7@cumc.columbia.edu (Z. Jin), whe@stats.uwo.ca (W. He).

http://dx.doi.org/10.1016/j.jmva.2016.02.006
0047-259X/

Lin [28]. They used a marginal proportional hazards model by modeling the effect of covariates with an unknown smooth nonparametric function. For this function, they developed kernel estimating equations by extending the Cox-type marginal estimating equations of Wei, Lin and Weissfeld [27], Lee, Wei and Amato [16] and Cai and Prentice [3], with the idea of kernel smoothing in Lin and Carroll [19]. The approach yields that the resulting asymptotic variance of the nonparametric estimator is minimized by assuming working independence rather than by assuming correct working correlation. It is also noted that the interpretation is based on the concept hazard which is a type of probability and cannot be used to quantify the change in survival time directly.

In this paper, we consider a natural extension of the classical nonparametric regression model for the analysis of clustered survival data, which models directly the effect of covariates on survival time with an unknown smooth nonparametric function; see the expression (1) in the next section for our model setup. Our estimation and inference approach is based on a class of unbiased data transformations and a local weighted least squares method. The key idea of the proposed method is to weight local observations with "local" variances, thus to improve the estimation efficiency. We derive the asymptotic properties of the resulting estimator, which show that the asymptotic variance of the nonparametric estimator is minimized with the correct specification of the correlation structure.

The remainder of this article is organized as follows. In Section 2, a nonparametric regression model framework is introduced. The estimating procedure and asymptotic properties are presented in Section 3. In Section 4, simulation studies with the proposed method are reported. In Section 5, the proposed method is illustrated with a real data from the Busselton Health Study. Section 6 contains concluding discussions. The sketch of proofs is provided in the Appendix.

2. Nonparametric regression model

Suppose that the data consist of n clusters with the ith cluster having J_i subjects, $i = 1, \ldots, n$. Let T_{ij} be the survival time and C_{ij} be the censoring time associated with the jth subject in the ith cluster. For the ith cluster, we have observations $\{(Y_{ij}, \delta_{ij}, X_{ij}) : j = 1, \ldots, J_i\}$, where $Y_{ij} = T_{ij} \wedge C_{ij}$, $\delta_{ij} = I\{T_{ij} \leq C_{ij}\}$ and X_{ij} is an individual level scalar covariate. It is assumed that T_{ij} and C_{ij} are conditionally independent given X_{ij}. The nonparametric regression model is specified as following:

$$T_{ij} = m(X_{ij}) + \varepsilon_{ij}, \tag{1}$$

where $m(\cdot)$ is an unknown smooth regression curve, ε_{ij} is the error term satisfying $\mathrm{E}(\varepsilon_{ij}) = 0$ with a finite variance. We assume that X_{ij} and ε_{ij} are independent. Write $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ij_i})^\top$, $\mathbf{T}_i = (T_{i1}, \ldots, T_{ij_i})^\top$, and $\mathbf{x}_i = (X_{i1}, \ldots, X_{ij_i})^\top$. Let $V_i = \mathrm{var}\{\mathbf{T}_i|\mathbf{x}_i\}$. We assume here that V_i could depend on an unknown parameter which could be consistently estimated separately from the data. The cluster sizes J_is are assumed to be bounded. To facilitate the presentation, we assume that $J_i \equiv J$ throughout the paper. It is assumed that $(\mathbf{T}_i, \mathbf{x}_i)$ are independent and identically distributed for $i = 1, \ldots, n$.

3. The estimation procedure

For censored data, however, $\mathbf{T}_i$s are not always observed, and thus the model setup (1) cannot be directly estimated. We propose to estimate $m(\cdot)$ based on a local linear regression with a class of unbiased data transformations being applied to the censored data. The use of unbiased data transformations is to adjust for the censoring effect in regression model (1). The estimation procedure involves two steps: an unbiased data transformation step and an estimation step for the parameters in the regression model. We will discuss and explain each step in the following two subsections.

3.1. The unbiased data transformation step

As considered in Zheng [29], Lai, Ying and Zheng [15], and Fan and Gijbels [6], we transform the triple $(Y_{ij}, \delta_{ij}, X_{ij})$ to (T^*_{ij}, X_{ij}) according to the rules,

$$T^*_{ij} = \delta_{ij}\phi_{1j}(X_{ij}, Y_{ij}) + (1 - \delta_{ij})\phi_{2j}(X_{ij}, Y_{ij}), \tag{2}$$

for $i = 1, \ldots, n, j = 1, \ldots, J$, where $\phi_{1j}(\cdot, \cdot)$ and $\phi_{2j}(\cdot, \cdot)$ are pre-specified known transformation functions. To replace T_{ij} with T^*_{ij} in model (1), a basic requirement for the transformation is to achieve unbiasedness, i.e., $\mathrm{E}(T^*_{ij}|X_{ij}) = \mathrm{E}(T_{ij}|X_{ij}) = m(X_{ij})$. Since $\phi_{1j}(\cdot, \cdot)$ and $\phi_{2j}(\cdot, \cdot)$ achieve this unbiasedness thus they are called unbiased data transformation functions. Specifically, the transformation functions $\phi_{1j}(\cdot, \cdot)$ and $\phi_{2j}(\cdot, \cdot)$ should satisfy the following equation for all x and y.

$$\int_0^\infty \left[\phi_{1j}(x, y)\bar{G}_j(y|x) - \int_0^y \phi_{2j}(x, c)\bar{G}_j(c|x)dc - y\right] d\bar{F}_j(y|x) = 0, \tag{3}$$

where $\bar{F}_j(y|x) = \Pr(T_{ij} > y|x)$ and $\bar{G}_j(y|x) = \Pr(C_{ij} > y|x)$ are the conditional survival functions for T_{ij} and C_{ij} given X_{ij}, respectively.

It is easy to see that the solution for (3) does not depend on the form of $\bar{F}_j(y|x)$ if

$$\phi_{1j}(x, y)\bar{G}_j(y|x) - \int_0^y \phi_{2j}(x, c)\bar{G}_j(c|x)dc - y = 0, \tag{4}$$