Computational Statistics and Data Analysis 87 (2015) 46–56

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A new estimating equation approach for marginal hazard ratio estimation

Yi Niu [a], Yingwei Peng [b,c,d,*]

[a] School of Mathematical Sciences, Dalian University of Technology, Dalian, Liaoning, 116024, China
[b] Department of Public Health Sciences, Queen's University, Kingston, ON, Canada K7L 3N6
[c] Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada K7L 3N6
[d] Cancer Care and Epidemiology, Queen's Cancer Research Institute, Kingston, ON, Canada K7L 3N6

ARTICLE INFO

Article history:
Received 7 March 2013
Received in revised form 25 November 2014
Accepted 21 January 2015
Available online 2 February 2015

Keywords:
Clustered failure times
Dependence structure
Estimating equation
Marginal model
Proportional hazards

ABSTRACT

Clustered failure time data often arise in biomedical studies and a marginal regression modeling approach is often preferred to avoid assumption on the dependence structure within clusters. A novel estimating equation approach is proposed based on a semiparametric marginal proportional hazards model to take the correlation within clusters into account. Different from the traditional marginal method for clustered failure time data, our method explicitly models the correlation structure within clusters by using a pre-specified working correlation matrix. The estimates from the proposed method are proved to be consistent and asymptotically normal. Simulation studies show that the proposed method is more efficient than the existing marginal methods. Finally, the model and the proposed method are applied to a kidney infections study.

1. Introduction

Clustered failure time data are frequently observed in biomedical and epidemiologic research. For example, in the Diabetic Retinopathy Study (Huster et al., 1989), each patient had one eye randomized to laser photocoagulation and the other eye received no treatment. The event of interest was the time from initiation of treatment to the time of occurrence of blindness. The times of blindness of two eyes from the same patients can be considered as clustered failure times with patients as clusters. In the study of infections of kidney patients (McGilchrist and Aisbett, 1991), several infections may be observed on each patient. The survival time was defined as the time between the end of the previous infection and the beginning of the current infection. The times to recurrence on each patient may be correlated to one another. Although it is reasonable to assume that the event times from the different subjects are independent, the correlation between the failure times from the same subject should not be ignored.

The proportional hazards model has been investigated extensively for multivariate failure time data with two most studied approaches including marginal models and frailty models. The marginal models provide population-averaged covariate effects while the dependence structure within a cluster is ignored. Examples include Huster et al. (1989), Wei et al. (1989), Lee et al. (1992), Liang et al. (1993), Lin (1994), Spiekerman and Lin (1998), Clegg et al. (1999), Yang and Ying (2001), Ebrahimi (2006), and Chen et al. (2010). On the other hand, when the association as well as the comparison of lifetimes within the same

* Corresponding author at: Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada K7L 3N6. Tel.: +1 613 533 6000x78525; fax: +1 613 533 6794.
E-mail address: pengp@queensu.ca (Y. Peng).

http://dx.doi.org/10.1016/j.csda.2015.01.012
0167-9473/

cluster is of interest, the frailty models have received considerable attention. The correlation structure in a frailty model is specified by a frailty or a random effect which is shared among observations in a cluster. The frailty variables from different clusters are usually assumed to be independent and to follow a parametric distribution such as the gamma distribution (Clayton, 1978; Clayton and Cuzick, 1985), the inverse Gaussian distribution (Hougaard, 1986a), the positive stable distribution (Hougaard, 1986b), and the lognormal distribution (McGilchrist and Aisbett, 1991; McGilchrist, 1993). Excellent discussions on the proportional hazards frailty model can be found in Hougaard (1995, 2000) and Therneau and Grambsch (2000).

To further improve the estimation efficiency of the marginal models, estimating equation approaches have been investigated by Segal and Neuhaus (1993), Cai and Prentice (1995, 1997), Prentice and Hsu (1997), and Gray and Li (2002), among others. This approach explicitly specifies working correlation structures in the estimating equations to accommodate the dependence of failure times in each cluster. Specifically, Segal and Neuhaus (1993) proposed a synthesis of the Poisson regression model and generalized estimating equations based on a parametric proportional hazards model for multivariate survival data. Cai and Prentice (1995, 1997) derived a weighted partial likelihood estimating equation based on a counting process approach for correlated failure time data. They developed the asymptotic distribution of the hazard ratio estimates from the proposed estimating equation. Prentice and Hsu (1997) extended Cai and Prentice (1997) by developing joint estimating equations for hazard ratio and pairwise dependence parameters. These joint estimating equations allow the correlation to depend on baseline covariates and on block size. Gray and Li (2002) considered the optimal selection of weights in martingale estimating equations for clustered failure time data based on the marginal proportional hazards model.

In this paper, we revisit the marginal method developed by Segal and Neuhaus (1993) for multivariate failure time data. We observe in numerical studies that, when correlation exists within clusters, the estimating function proposed by Segal and Neuhaus (1993) for hazards ratio parameters tends to be biased. Therefore, the estimates from the existing estimating equation have large biases and the variance estimates are unstable. To address this issue, we propose an unbiased weighted estimating function and show that the estimators based on the proposed estimating equation are consistent and asymptotically normal. A consistent estimator of the covariance matrix for regression parameters is also provided. We will demonstrate via a simulation study that the proposed estimating equation produces unbiased regression estimators as well as improves the estimation efficiency compared to the existing marginal methods.

The rest of the paper is organized as follows. In Section 2, we propose an unbiased weighted estimating function for the hazard ratio parameters based on the marginal semiparametric proportional hazards model. The asymptotic properties of the estimators of the hazard ratio parameters are presented in this section. We perform a simulation study in Section 3 to evaluate the performance of the proposed estimating equation, and apply this approach to a study of infections of kidney patients in Section 4. Finally, we provide conclusions on the proposed model and estimation methods in Section 5.

2. Marginal model with correlation structures

Let $\tilde{T}_{ij}$ and C_{ij} be the failure and censoring times for the jth subject in the ith cluster ($j = 1, \ldots, n_i, i = 1, \ldots, K$), and $N = \sum_{i=1}^{K} n_i$. The observed failure time is $T_{ij} = \min(\tilde{T}_{ij}, C_{ij})$, its censoring status is denoted as $\delta_{ij} = I(\tilde{T}_{ij} \le C_{ij})$ ($I(A) = 1$ if A is true and 0 otherwise), X_{ij} is a vector of p_X covariates that may have effects on the failure time distribution. We assume that C_{ij} is independent of $\tilde{T}_{ij}$ given X_{ij}.

The marginal survival function of $\tilde{T}_{ij}$ is assumed to follow the proportional hazards model, i.e.,

$$S(t; X_{ij}) = S_0(t)^{\exp(\beta' X_{ij})}, \tag{1}$$

where $S_0(t)$ is a baseline survival function of $\tilde{T}_{ij}$ when $X_{ij} = 0$, and β is a vector of $p_X \times 1$ unknown hazard ratio parameters for X_{ij}. Following Kalbfleisch and Prentice (2002), we let $\tau_1 < \cdots < \tau_k$ be the distinct uncensored failure times, and $S_0(t) = 1$ if $t < \tau_1$, otherwise $S_0(t) = \prod_{s:\tau_s \le t} \exp(-\alpha_s)$, where $\alpha = (\alpha_1, \ldots, \alpha_k)$ are k nonnegative parameters. Let $\theta = (\beta, \alpha)$. We further assume that given X_{ij}, $\tilde{T}_{ij}$ and $\tilde{T}_{ij'}$ are correlated due to the shared environment or genetic in a cluster if $j \ne j'$. However, $\tilde{T}_{ij}$ and $\tilde{T}_{i'j'}$ are independent if $i \ne i'$.

If we ignore the correlation within clusters, the unknown parameters in the marginal model (1) could be estimated based on a log-likelihood function with the observations $O = \{(T_{ij}, \delta_{ij}, X_{ij}), j = 1, \ldots, n_i, i = 1, \ldots K\}$. That is,

$$\begin{aligned} l(\theta; O) &= \log \prod_{i=1}^{K} \prod_{j=1}^{n_i} h(t_{ij}; X_{ij})^{\delta_{ij}} S(t_{ij}; X_{ij}) \\ &= \log \prod_{i=1}^{K} \prod_{j=1}^{n_i} \left[\{\exp(\beta' X_{ij})\}^{\delta_{ij}/\Lambda_0(t_{ij})} \exp\{-\exp(\beta' X_{ij})\}\right]^{\Lambda_0(t_{ij})} + \log \prod_{i=1}^{K} \prod_{j=1}^{n_i} \left(\lambda_0(t_{ij})\right)^{\delta_{ij}}, \end{aligned} \tag{2}$$

where $h(t_{ij}; X_{ij})$ is the hazard function corresponding to $S(t_{ij}; X_{ij})$, and $\lambda_0(t_{ij})$ and $\Lambda_0(t_{ij})$ are the hazard and cumulative hazard functions corresponding to $S_0(t_{ij})$.