Computational Statistics and Data Analysis 103 (2016) 79–90

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Cox regression analysis of dependent interval-censored failure time data

Ling Ma [a,*], Tao Hu [b], Jianguo Sun [a]

[a] *Department of Statistics, University of Missouri, Columbia, MO, United States*

[b] *School of Mathematical Sciences and BCMIIS, Capital Normal University, Beijing, China*

ARTICLE INFO

Article history:
Received 28 July 2015
Received in revised form 21 April 2016
Accepted 27 April 2016
Available online 11 May 2016

Keywords:
Copula model
Cox model
Dependent interval-censoring
Efficient semiparametric estimation

ABSTRACT

Many procedures have been proposed for regression analysis of interval-censored failure time data arising from the Cox or proportional hazards model. However, most of these existing methods only apply to the situation where the censoring mechanism generating censoring intervals is independent of the failure time of interest, and it is well-known that sometimes this may not be true in practice. To address this issue, a new approach that allows the dependence between the censoring mechanism and the failure time is proposed. More specifically, a situation where the dependence is through the length of censoring intervals is considered as it is often the case in follow-up studies. The asymptotic properties of the proposed estimators are established and the numerical studies are conducted for the assessment of the finite sample properties of the estimators.

1. Introduction

This paper discusses regression analysis of interval-censored failure time data arising from the Cox or proportional hazards model. By interval-censored data, we mean that instead of being observed exactly, the failure time of interest is observed only to belong to an interval, commonly denoted by $(L, R]$ or $L < T \leq R$ (Chen et al., 2012; Finkelstein, 1986; Huang, 1996; Sun, 2006). That is, the occurrence of the failure event of interest is known only to be within the interval $(L, R]$. Examples of the areas that often produce such data include health or medical follow-up studies such as clinical trials as well as social sciences. In addition, it is easy to see that the interval-censored data include right-censored data, whose analysis has been extensively discussed in the literature, as a special case (Kalbfleisch and Prentice, 2002).

The analysis of interval-censored failure time data has been attracting more and more attention recently and especially, many methods have been developed for their regression analysis under various regression models including the proportional hazards model. For example, a pioneering work was given by Finkelstein (1986), which discussed the fitting of the Cox model to the data, and for the same problem, Goggins et al. (1998) and Zhang et al. (2010) developed a Markov chain Monte Carlo EM algorithm and a spline-based maximum likelihood approach, respectively. Also among others, Zhang et al. (2005) proposed an estimating equation procedure for fitting the linear transformation model to interval-censored data. More and relatively complete references on this can be found in Chen et al. (2012) and Sun (2006).

For the analysis of interval-censored failure time data, it is easy to see that in addition to the failure time T of interest, one may need to concern the mechanism that generates L and R also. For this, most of the existing methods assume the noninformative or independent mechanism under which one can carry out the conditional analysis given $L < T \leq R$.

* Corresponding author.
E-mail address: mlbegood@gmail.com (L. Ma).

http://dx.doi.org/10.1016/j.csda.2016.04.011
0167-9473/

On the other hand, it is apparent that this may not be true in some situations. In a clinical trial with the follow-up of both healthy subjects and patients, for example, the patients may tend to pay more clinical visits than the scheduled ones. Several methods have been proposed in the literature for the analysis of dependent interval-censored data. For example, Finkelstein et al. (2002) and Betensky and Finkelstein (2002) considered the one-sample problem related to such data, and Zhang et al. (2007) gave a latent variable-based estimation procedure for regression analysis of dependent interval-censored data from the Cox model. Note that the method in Zhang et al. (2007) used a log-normal frailty to describe the dependence structure and it is apparent that in reality, it may be difficult to check such specific dependence structure. In other words, a more general model or procedure would be clearly needed.

In the following, we will discuss the same problem considered as Zhang et al. (2007) and present a more general, copula model-based approach. More specifically, we will assume that the pair $(L, R]$ is generated from an underlying observation process and focus on the case where the dependence between the failure time and the censoring mechanism can be described by the length of censoring intervals. One situation where this can happen is medical follow-up studies where subjects may pay more or less clinical visits. For the analysis, a maximum likelihood estimation procedure will be developed with the use of I-spline functions (Ramsay, 1988; Lu et al., 2007). Both the model and the method will be described in the next section. In Section 3, some asymptotic properties of the proposed estimators are established, including the consistency and asymptotic normality of the estimated regression parameters, and the variance estimation is discussed. Some simulation results are presented in Section 4 and Section 5 provides an illustrative example. Finally Section 6 contains some discussion and concluding remarks.

2. Semiparametric maximum likelihood estimation

Consider a failure time study that involves n independent subjects and yields only interval-censored failure time data. For subject i, let T_i denote the failure time of interest and Z_i the vector of covariates, and suppose that there exists an array of underlying observation times $\boldsymbol{E}_i = \{E_{i,j} : j = 0, 1, 2, \ldots\}$ such that $E_{i,j} < E_{i,j'}$ for any $j < j'$. Define $W_{i,j} = E_{i,j} - E_{i,j-1}$, $j = 1, 2, \ldots$, the gap times of the observation process with $E_{i0} = 0$ and assume that given Z_i, the $W_{i,j}$'s follow the same distribution. Furthermore assume that T_i may depend on the underlying observation process through the gap times and that given Z_i, $\{(T_i, W_{i,j}),\ j = 1, 2, \ldots\}$ follow the same joint distribution. In the following, for each i, define $(\tilde{L}_i, \tilde{R}_i]$ to be the random interval among $(0, E_{i,1}]$, $(E_{i,1}, E_{i,2}]$, ... that contains T_i. Note that the mechanism behind the observation process and censoring intervals here is similar to that behind the mixed case k interval-censored data (Schick and Yu, 2000; Sun, 2006).

In practice such as in medical follow-up studies, there usually exists an administrative censoring time ζ_i beyond which the observation process is no longer available. For each i, define $L_i = \max\{E_{i,j} : E_{i,j} < \min(\zeta_i, T_i),\ j = 0, 1, 2, \ldots\}$ and $R_i = \min\{E_{i,j} : E_{i,j} > L_i,\ j = 1, 2, \ldots\}$. Also, define $\delta_i = 1$ if $R_i \le \zeta_i$ and $\delta_i = 0$ otherwise. Note that when $\delta_i = 0$, R_i is not observed but right-censored at ζ_i and T_i is right-censored. For ease of explanation, we define $(\tilde{L}_i, \tilde{R}_i]$ as covering interval and $(L_i, R_i]$ as 'observed interval' (even though R_i is not directly observed for right-censored subjects). And define $W_i = R_i - L_i$ as the 'observed interval length'. It is easy to see that if $\delta_i = 1$, W_i is exactly observed and T_i lies in the observed interval. For subjects with $\delta_i = 0$, note that L_i is not necessarily equal to $\tilde{L}_i$ and also $(L_i, R_i]$ may not contain T_i. In fact, since R_i is right-censored at ζ_i, we only observe that $T_i > L_i$ and $W_i > \zeta_i - L_i$. An illustrative example for this situation with $\delta_i = 0$ is given in Fig. 1. Assume that given Z_i and T_i belonging to the censoring interval, the distribution of T_i depends on the censoring interval only through its length. That is, we have

$$Pr(T_i \le t \mid L_i < T_i \le R_i; L_i = l_i, R_i = r_i, Z_i = z_i) = Pr(T_i \le t \mid l_i < T_i \le r_i; W_i \equiv r_i - l_i, Z_i = z_i).$$

Then the likelihood function has the form

$$\prod_{i=1}^{n} \left[Pr(L_i = l_i < T_i \le R_i = r_i, W_i = w_i)\right]^{\delta_i} \left[Pr(T_i > L_i = l_i, W_i > \zeta_i - l_i)\right]^{1-\delta_i}.$$

Let K denote the joint distribution of T and W given Z. Then it is well-known (Nelsen, 2006) that there exists a copula function $C_\alpha(u, v)$ defined on $I^2 = [0, 1] \times [0, 1]$ such that

$$K(t, w) = C_\alpha(F_T(t), F_W(w)), \quad t \ge 0,\ w \ge 0.$$

In the above, F_T and F_W denote the marginal distributions of the T and W given Z, respectively, α, usually referred to as the association parameter, represents the relationship between T and W, and $C_\alpha(u, 0) = C_\alpha(0, v) = 0$, $C_\alpha(u, 1) = u$ and $C_\alpha(1, v) = v$. Define $m_\alpha(F_T(t), F_W(w)) = P(T \le t | W = w, Z)$. Then we have

$$m_\alpha(F_T(t), F_W(w)) = \left.\frac{\partial C_\alpha(u, v)}{\partial v}\right|_{u=F_T(t), v=F_W(w)}.$$

For the covariate effects, we will consider the following marginal Cox hazard models

$$\lambda_T(t|Z_i) = \lambda_1(t)\exp(Z_i'\beta), \qquad \lambda_W(w|Z_i) = \lambda_2(w)\exp(Z_i'\gamma),$$

for T_i and W_i given Z_i, respectively. Here $\lambda_1(t)$ and $\lambda_2(w)$ are unknown baseline hazard functions and β and γ denote vectors of regression parameters.