Regression analysis of doubly censored failure time data with frailty

Yang-Jin Kim [1]

ABSTRACT

The timings of two successive events of interest may not be measurable, instead it may be right censored or interval censored; this data structure is called doubly censored data. In the study of HIV, two such events are the infection with HIV and the onset of AIDS. These data have been analyzed by authors under the assumption that infection time and induction time are independent. This paper investigates the regression problem when two events are modeled to allow the presence of a possible relation between two events as well as a subject-specific effect. We derive the estimation procedure based on Goetghebeur and Ryan's (2000) piecewise exponential model and Gauss-Hermite integration is applied in the EM algorithm. Simulation studies are performed to investigate the small-sample properties and the method is applied to a set of doubly censored data from an AIDS cohort study.

Keywords: Doubly censored data; Frailty; Regression analysis; EM algorithm; AIDS cohort study.

[1] Senior Researcher, Institute of Statistics, Korea University, Seoul 136-701, Korea

1. Introduction

Doubly censored data arise when the times of occurrence of two successive events are not directly observable, instead intervals that include these times are available. An example of doubly censored data occurs in acquired immune deficiency syndrome (AIDS) cohort studies. In this situation, main interests are the estimation of the distribution of the induction time, which is defined as the time between infection with Human immunodeficiencies virus (HIV) and the onset of AIDS, and the effects of covariates on the induction time.

De Gruttola and Lagakos (1989) proposed a method to estimate the distribution of infection time and induction time by using self-consistency algorithm of Turnbull (1976); Gomez and Lagakos (1994) developed a two-stage estimation procedure; Sun (1995, 1997) considered the same problems for truncated doubly censored data. For regression analysis, Kim et al. (1993) considered the maximum likelihood method under the assumption that infection time and induction time are discrete random variables and estimated regression parameters as well as distributions using the EM algorithm. Goggins et al. (1999) presented MCEM approach to estimate regression parameters without considering the distribution of infection time. Sun et al. (1999) derived the simple estimating equation, and Pan (2001) used the multiple imputation technique to solve this estimating equation. Most methods assume that induction time follows the proportional hazard model. Recently, Sun et al. (2003) used the additive hazards model.

In all of these methods, it is assumed that infection time and induction time are independent. In this paper, we consider a methodology to estimate the distributions of infection time, induction time, and the effects of covariates on induction time when this assumption is questionable. A frailty effect is employed to both infection time and induction time to represent common factors as well as heterogeneities. A frailty model is widely used to incorporate the dependency between subjects in the same cluster and has been studied by authors including Clayton and Cuzick (1985), Klein (1992), and Nielsen et al. (1992). In the right censored data, a maximum likelihood estimator in the semiparametric frailty model was obtained by using the EM algorithm with Gamma distributed frailties. Hougaard (1986) considered a multivariate failure time model, in which a frailty was assumed to follow a positive stable distribution. Vaida and Xu (2000) and Maples et al. (2002) derived the estimation procedure by using the MCEM algorithm when random effects followed normal distribution. Recently, to overcome several disadvantage of univariate frailty into multivariate survival time, bivariate frailty model has been widely used. Xue and Brookmeyer (1995) proposed a bivariate log normal frailties to characterize association for multivariate survival times and developed a modified EM algorithm. Ng and Cook (1998) showed the application of bivariate frailties to two state model.

In this paper, we employ a bivariate frailty model to assign different frailty to each event and to incorporate a association by allowing an association under doubly censored data. For estimation, we use the EM algorithm and perform a sampling in E-step to recover two kinds of missing data caused by interval censoring and frailty term. For derivation of estimates, our approach is based on Goetghebeur and Ryan's (2000) piecewise exponential model. Their method assumes that the unit of time is chosen finely enough for each event time to be rounded up to the endpoint of the interval within which it occurred. We extend this model to doubly censored data and introduce frailty effect to incorporate both a subject-specific effect and a possible relation between two events.

The remainder of this paper is organized as follows: Section 2 introduces the notation and model. In Section 3, and the application to data from an AIDS follow-up study appears in Section 2.

2. Statistical Model

Consider a study that involves n independent subjects and in which each subject experiences two successive events. For example, in the AIDS study, C_i denotes the time of HIV infection, and Z_i denotes the time of AIDS onset of subject $i, i = 1, \ldots, n$. Now, the induction time T_i is defined as $T_i = Z_i - C_i$. In doubly censored data, both C_i and Z_i are not directly observed and, instead the available data are the intervals $[C_{Li}, C_{Ri}]$ and $[Z_{Li}, Z_{Ri}]$. Therefore, induction time is also interval censored by $[T_{Li}, T_{Ri}]$, where $T_{Li} = Z_{Li} - C_{Ri}$ and $T_{Ri} = Z_{Ri} - C_{Li}$.

Kim et al. (1993) assumed that C_i and T_i are discrete random variables and derive full likelihood using indicators that show admissible values of (C, T). They used the self-consistency algorithm and Newton-Raphson algorithm to estimate the distribution of infection time, induction time, and regression parameters. Sun et al. (1999) derived an estimating equation to estimate regression parameters; Pan (2001) also considered similar method and multiple imputation was used to estimate the distribution of infection and regression parameters.

To consider a possible relation between infection time and induction time and a subject-specific effect, we introduce the bivariate frailty effects, $u = (u_1, u_2)$. Specifically, given $u_i = (u_{i1}, u_{i2})$, the intensities of C and T of a subject i are assumed to be given by

$$\lambda^C(c \mid u_{i1}) = \exp(u_{i1})\ \tilde{\lambda}(c), \tag{1}$$

$$\lambda^T(t \mid u_{i2}, X_i) = \exp(\beta' X_i + u_{i2})\lambda(t), \tag{2}$$

where $\tilde{\lambda}(c)$ is an unknown baseline hazard function of infection time, $\lambda(t)$ is an unknown baseline hazard function of induction time, X_i represents a vector of covariates, and β denotes the corresponding regression parameters. We assume that infection time and induction time are independent given u_i. To accommodate both positive and negative correlation between two variables, we consider the bivariate normal distribution. We therefore assume that $u_i = (u_{i1}, u_{i2})' \sim N(0, \Sigma), i = 1, \ldots, n$, where

$$\Sigma = \begin{pmatrix} \sigma_1, & \sigma_3 \\ \sigma_3, & \sigma_2 \end{pmatrix}$$

3. AIDS Cohort Study

We analyzed the AIDS cohort study from Kim et al.(1993), where two covariates are available, treatment and age. The main interest is the comparison of two treatments with respect to the induction time. Two treatment groups are divided based on the amount of blood factor a patient received into lightly treated or heavily treated group. Of th 257 hemophiliacs involved in the study since 1978, 153 have been lightly treated and 104 patients are heavily treated. At the end of this study, 188 patients were found to be HIV positive, and 41 had developed to either AIDS, ARC (Aids related complex) or a low platelet count. Define $X_1 = 0$, if a patient is in a lightly treated group and $X_1 = 1$ if a patient is in a heavily treated group. To compare our result to others' studies, Table1 shows estimates and their corresponding standard errors. For bivariate normal frailties, our point estimates are ($\hat{\beta} = 0.67$), with standard error is 0.33. $(\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3) = (0.307, 0.774, 0.106)$ and it makes the treatment effect significant ($p = 0.021$). When apply Gamma frailty, we have ($\hat{\beta}_G = 0.65$) with standard error, 0.41 and $\hat{\alpha} = 0.787$. Kim et al. (1993) also considered the age effect on induction time where $X_2 = 0$ if subject's estimated infection age is younger than 20 years, and $X_2 = 1$ otherwise. Sun et al. (1999) also analyzed the effects of treatment and age and they gave $\tilde{\beta}_1$=0.71 and $\tilde{\beta}_2$=0.053 with estimated standard errors 0.29 and 0.35, respectively. When bivariate frailties model is applied, we have $\hat{\beta}_1 = 0.712$ and $\hat{\beta}_2 = 0.107$ for treatment and age effect with estimated standard errors 0.33 and 0.34. Also for estimates of covariance, $(\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3)$=(0.180, 0.883, 0.0346). For gamma frailty model, $\hat{\beta}_{G1}$=0.69 and $\hat{\beta}_{G2}$=0.105; estimated standard errors 0.42 and 0.30, respectively, and $\hat{\alpha}$=0.763. Figure 1 shows estimated

Gamma frailty effects of 188 patients. We observe several patients have frailty values which are far from 1. Also, most of patients with interval censored induction time have bigger frailty values than these with right censored data.

Table 1. Treatment effect for AIDS induction time

Model	Estimate	Standard Error
Kim et al.(1993)	0.69	0.34
Goggins et al.(1999)	0.68	0.34
Pan (2001)	0.71	0.33
Sun et al. (1999)	0.70	0.28
Midpoint imputation	0.74	0.33
Bivariate frailties	0.67	0.33
Gamma frailty	0.65	0.41

REFERENCES

Goetghebeur, E. and Ryan, L. (2000). Semiparametric Regression Analysis of Interval-Censored Data. *Biometrics* **56,** 1139-1144.

Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* **48**, 795-806.

Lindsey, J. and Ryan, L. (1998). Methods for interval censored data. Tutorial in biostatistics. *Statistics in Medicine* **17,** 219-238.

Sun, L. , Kim, Y., and Sun, J. (2003). Regression analysis of doubly censored failure time data using additive hazard model. *Technical Report*, Department of Statistics, University of Missouri.