Regression Analysis of Doubly censored data using Gibbs Sampler for the Incubation period

Hanna Yoo[1] Jae Won Lee[2]

Abstract

In standard time-to-event or survival analysis, the occurrence times of the event of interest are observed exactly or are right-censored. However in certain situations such as the AIDS data, the incubation period which is the time between HIV infection time and the diagnosis of AIDS is usually doubly censored. That is the HIV infection time is interval censored and also the time of the diagnosis of AIDS is right censored. In this paper, we impute the interval censored infection time using the conditional mean imputation and estimate the coefficient factor of the regression analysis for the incubation period using Gibbs sampler. We applied parametric and semi-parametric methods for the analysis of the Incubation period and compared the results.

Keywords : Doubly censored data, conditional mean imputation, Gibbs sampler

1. Introduction

In standard time-to-event or survival analysis, the occurrence times of the event of interest are observed exactly or are right-censored. However in certain situations such as the AIDS data, the incubation period which is the time between HIV infection time and the diagnosis of AIDS is usually doubly censored. That is the HIV infection time is interval censored and also the time of the diagnosis of AIDS is right censored. Many statistical approaches have been applied to estimate the incubation period distribution and also find out the relationships between the covariates. Many methods for analyzing doubly censored data have been proposed(Victor De Gruttola et al., 1989, Mimi Y.Kim et al., 1993, Jianguo Sun., 1995, Wei Pan., 2001). All these methods are based on frequentist paradigm when estimating the covariate effect on the incubation period. They also use the origin doubly censored data which is hard to analyze with the existing statistical program such as SAS or S-Plus. In this paper we imputed HIV infection time using the conditional mean imputation method and changed the doubly censored data to right censored data and applied Bayesian paradigm when estimating the covariate effect using Gibbs Sampler.

In our study we applied parametric and a semi-parametirc model to the survival times and used Gibbs sampler to calculate the posterior distribution of the covariate

1) Graduate student, Department of Statistics, Korea University, Seoul 136-701
2) Professor, Department of Statistics, Korea University, Seoul 136-701

effect.

2. Conditional mean Imputation

In the AIDS study where the data are doubly censored it is hard to calculate the incubation time. Many Imputation methods are used to impute the HIV infection time. There are single imputation methods such as "midpoint imputation" ," mean imputation", "hot deck" , etc and multiple imputation such as "Approximate Bayesian Bootstrap" , "Poor Man's Data Augmentation", etc. There are advantages and disadvantages in both methods. In this paper we applied a single imputation method, conditional mean imputation which is rather easy to implement than the multiple imputation method.

In some cohort studies where individuals are periodically screened for the evidence of infection the exact infection time are thought to be in the interval (L_i, R_i) where L_i is the known calender time of the last negative test and R_i is the known calender time of the first positive test. The simplest method estimating the infection time is using the midpoint of the interval but this method has relatively high bias. Conditional mean imputation is a method using the expected infection data given that infection time occurred between L_i and R_i. Let f be the pdf of infection times among individuals and F be the following CDF. Then the imputed infection time is estimated as:

$$\hat{s}_i = \frac{\int_{Li}^{R_i} s\hat{f}(s)ds}{\hat{F}(R_i) - \hat{F}(L_i)}, \text{ where } \hat{F}(t) = \int_0^t \hat{f}(s)ds$$

To get the imputed infection time we need a parametric model for the infection time and in our study the weibull model was selected among many other parametric models and the whole data has set up to a right censored data.

3. Gibbs Sampler

The Gibbs sampler which is used through out this paper when estimating the covariate effect of the incubation period distribution, is one of the best known MCMC sampling algorithms in the Bayesian computation literature which is founded on the ideas of Grenander(1983) and the formal term is introduced by Geman and Geman(1984). It evaluates high-dimensional posterior integrals and is easy to implement. Let $\theta = (\theta_1, \theta_2, .., \theta_p)'$ be a p-dimensional vector of parameters and let $\pi(\theta|D)$ be its posterior distribution given the data D. The basic scheme of the Gibbs sampler is given as follows:

Step1: Choose an arbitrary staring set of values $\theta_0 = (\theta_{1,0}, \theta_{2,0}, \cdots, \theta_{p,0})'$ and set $i = 0$.
Step2: Generate $\theta_{i+1} = (\theta_{1,i+1}, \theta_{2,i+1}, \cdots, \theta_{p,i+1})'$ as follows:

- Generate $\theta_{1,i+1} \sim \pi(\theta_1|\theta_{2,i},\ldots,\theta_{p,i},D)'$;
- Generate $\theta_{2,i+1} \sim \pi(\theta_2|\theta_{1,i+1},\theta_{3,i},\ldots,\theta_{p,i},D)'$;

... ...

- Generate $\theta_{p,i+1} \sim \pi(\theta_p|\theta_{1,i+1},\theta_{2,i+1},\ldots,\theta_{p-1,i+1},D)'$;

Step3: Set $i=i+1$, and go the step2.

It is shown that under certain regularity conditions, the vector sequence $\{\theta_i, i=1,2,\ldots\}$ has a stationary distribution $\pi(\theta|D)$.

3.1 Parametric Approach

To estimate the covariate effect of the incubation period distribution we assumed the incubation period distribution following a weibull distribution which is the most widely used parametric survival model. Suppose we have independent identically distributed survival times $y=(y_1,\ldots,y_n)'$, each having a Weibull distribution with the pdf:

$$f(y|\alpha,\lambda)=\alpha y^{\alpha-1}\exp(\lambda-\exp(\lambda)y^{\alpha}).$$

To build the Weibull regression model, we introduce the covariate through $\lambda_i=x_i'\beta$. Then the joint posterior has the form as follows:

$$\pi(\beta,\alpha|D)\propto \alpha^{\alpha_0+d-1}\exp\{\sum_{i=1}^{n}(\nu_i x_i'\beta+\nu_i(\alpha-1)\log(y_i)-y_i^{\alpha}\exp(x_i'\beta)) - \kappa_0\alpha-\frac{1}{2}(\beta-\mu_0)\sum^{-1}(\beta-\mu_0)\}$$

where α has a gamma prior $G(\alpha_0,\kappa_0)$ and a normal prior $N(\mu_0,\sigma_0^2)$ for λ.
For the Weibull regression model the conditional posterior distribution of $[\alpha|\beta,D]$ and $[\beta|\alpha,D]$ are log-concave and so implementation of the Gibbs sampler is straightforward.

3.2 Semi-Parametric Approach

We applied another approach through semi-parametric modeling. The gamma process is perhaps the most commonly used nonparametric prior process for the baseline cumulative hazard function in the Cox model. The process $\{Z(t): t \geq 0\}$ is called gamma process and is denoted by $Z(t) \sim GP(c\alpha(t),c)$ when it has the properties as follows:

$(i) Z(0) = 0$

$(ii) Z(t)$ has independent increment disjoint intervals

$(iii) t > s, Z(t) - Z(s) \sim G(c(\alpha(t) - \alpha(s)), c)$ where c is the weight about the mean $\alpha(t)$.

When modeling the semi-parametric we use the counting process notation. For subjects $i = 1, \ldots, n$, we observe processes $N_i(t)$ which count the number of failures at time t. The intensity process $I_i(t)$ is given by

$$I_i(t)dt = E(dN_i(t) \mid F_{t-})$$

where $dN_i(t)$ is the increment of N over [t , t+ dt) and F_{t-} is the available data just before time t. As $dt \rightarrow 0$ this becomes the instantaneous hazard at time t for subject i. This is assumed to have the proportional hazards form

$$I_i(t) = Y_i(t)\lambda_0(t)\exp(\beta' z_i)$$

The joint posterior distribution is defined by

$$P(\beta, \Lambda_0() \mid D) \propto P(D \mid \beta, \Lambda_0()) P(\beta) P(\Lambda_0())$$

where $D = \{ N_i(t), Y_i(t), z_i; i = 1, \ldots, n \}$ is the observed data. For implementing the Gibbs sampler variables are assigned prior distributions as follows:

$$dN_i(t) \sim Poisson(I_i(t)d(t))$$
$$I_i(t)d(t) = Y_i(t)\exp(\beta' z_i)d\Lambda_0() \quad (d\Lambda_0() = \lambda_0(t)dt)$$
$$d\Lambda_0(t) \sim G(cd\Lambda_0^*(t), c)$$
$$\Lambda_0(t) \sim GP(cd\Lambda_0(t))$$
$$\Lambda_0^*(t) = rdt$$

where r is the prior guess at the failure rate per unit time and c is a weak prior belief.

4. Results

To implement our analysis we applied real data discussed in Kim et.al (1993) of

the AIDS cohort study of hemophiliacs. This study consists of individuals with Type A or B hemophilia who were at risk fo HIV infection through the contaminated blood factor they received for their treatment. The subjects were classified into two groups according to the amount of blood they received. The data have 188 HIV infected subjects with the infection time interval censored and the AIDS onset time right censored. We imputed the HIV infection time with conditional mean imputation method and applied parametric and a semi-parametirc model to the incubation period and estimated the covariate effect through Gibbs sampler. We chose several different prior distributions and compared the results. As a result the coefficient factor of the regression analysis for the incubation period suggest that the subject in the heavily treated group had shorter incubation period then the lightly treated group.

References

[1] Victor De Gruttola, Stephan W. Lagakos(1989), Analysis of Doubly-Censored Survival Data, with Application to AIDS, Biometrics, Volume 45, Issue 1 (Mar., 1989), 1-11.

[2] Mimi Y. Kim Victor G. De Gruttola, Stephen W.Lagakos(1993), Analyzing Doubly Censored Data with Covariates, with Application to AIDS, Biometrics, Volume 49, Issue 1 Mar., 1993), 13-22.

[3] Debajyoti Sinha, Dipak K. Dey(1997), Semiparametric Bayesian Analysis of Survival Data, JASA, Vol.92, No. 439, Review paper.

[4] Jianguo Sun, Qiming Liao, and Marcello Pagano(1999), Regression Analysis of Doubly Censored Failure Time Data with Applications to AIDS Studies, Biometrics, Volume 35, 909-914.

[5] Joseph G. Ibrahim, Ming-Hui Chen, Debajyoti Sinha(2001), Bayesian Survival Analysis, Springer, America.