The Korean Journal of Applied Statistics (2009)
22(3), 607–616

Comparing Imputation Methods for Doubly Censored Data

Hanna Yoo[1] · Jae Won Lee[2]

[1]Department of Statistics, Korea University; [2]Department of Statistics, Korea University

(Received January 2009; accepted April 2009)

Abstract

In many epidemiological studies, the occurrence times of the event of interest are right-censored or interval censored. In certain situations such as the AIDS data, however, the incubation period which is the time between HIV infection and the diagnosis of AIDS is usually doubly censored. In this paper, we impute the interval censored HIV infection time using three imputation methods. Mid imputation, conditional mean imputation and approximate Bayesian bootstrap are implemented to obtain right censored data, and then Gibbs sampler is used to estimate the coefficient factor of the incubation period. By using Bayesian approach, flexible modeling and the use of prior information is available. We applied both parametric and semi-parametric methods for estimating the effect of the covariate and compared the imputation results incorporating prior information for the covariate effects.

Keywords: Doubly censored data, conditional mean imputation, approximate Bayesian bootstrap, Gibbs sampler.

1. Introduction

Doubly censored data can be found in many epidemiological studies especially in the AIDS data. The incubation period which is the time between HIV infection 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 doubly censored incubation period distribution and to find out the relationships with the covariates. Sun (2004) reviews the statistical methods that have been proposed in many literatures for analyzing doubly interval censored data. Following Sun (2004), most researchers focus on two basic problems: nonparametric estimation and regression analysis of the AIDS incubation period. For the nonparametric estimation, several authors have proposed to estimate the distribution function of the AIDS incubation time (de Gruttola and Lagakos, 1989; Gomez and Calle, 1999; Sun, 1995). For the regression analysis, Kim *et al.* (1993) used the full likelihood to analyze AIDS incubation time and Sun *et al.* (1999) proposed an estimating equation based method to investigate the effect of

This work was supported by the Korea Research Foundation Grant funded by the Korean Government(KRF-2008-313-C00148).

[2]Corresponding author: Professor, Department of Statistics, Korea University, Seoul 136-701, Korea.
E-mail: jael@korea.ac.kr

covariates on AIDS incubation time. These regression methods are the semi-parametric inferences based on the proportional hazards model. In addition the two-stage parametric model for estimating the effect of covariates of AIDS incubation time for doubly censored data has been proposed by Brookmeyer and Goedert (1989).

All these proposed methods for nonparametirc, semi-parametric and parametric based inference use the original doubly censored data which are hard to analyze. There was a modification in using the original doubly censored data and many researchers applied the imputation methods. Law and Brookmeyer (1992) studied the effect of mid-point imputation on the analysis of doubly censored data and Pan (2001) proposed a multiple imputation approach for Cox regression analysis to both interval and doubly censored data. Geskus (2001) considered four different imputation methods to impute the interval censored HIV infection time and estimated the doubly censored AIDS incubation period. With or without imputation, most of the papers are based on the frequentist paradigm when estimating the covariate effect on the incubation period. In this paper we propose to use Bayesian approach when estimating the covariate effect.

Recently Bayesian analysis have been applied to survival data more frequently due to its advanced computational skills. Generally nonparametric Bayesian methods (Kalbfleisch, 1978; Arjas and Gasbarra, 1994; Berliner and Hill, 1988) and semi-parametric methods (Burridge, 1981; Sinha and Dey, 1997) are applied. Bayesian paradigm has advantage for modeling complex survival models due to censoring and it enables to incorporate prior information. However, the use of Bayesian approaches with doubly censored data has been rather limited especially with the AIDS study. This is due to the short history of AIDS research and thus there does not exist much information on the distribution of the AIDS incubation period.

In this paper we compare three imputation methods when estimating the covariate effect through Bayesian approach. We consider mid imputation and the conditional mean imputation both of which are single imputation methods and also the approximate Bayesian bootstrap(ABB) which is a multiple imputation method. These imputation methods change the doubly censored AIDS incubation period data to right censored data and make it easy to analyze. After imputing the interval censored HIV infection time, Bayesian approach is used to estimate the covariate effect.

2. Background

2.1. Imputation methods

In the AIDS study where the data are doubly censored, it is hard to calculate the incubation period and many imputation methods are used to impute the HIV infection time. There are single imputation methods such as "midpoint imputation", "mean imputation", "conditional mean imputation", "hot deck" *etc.*, and multiple imputation such as "approximate Bayesian bootstrap", "poor man's data augmentation" *etc.*. There are advantages and disadvantages of both methods. For the single imputation method, it is easy to implement but they do not account for variability between imputations. On the other hand, multiple imputation minimizes standard error and increases efficiency of the estimates but they are more difficult to perform (Rubin,1987). In this paper, we consider three general imputation methods for missing data problems: mid imputation, conditional mean imputation and approximate Bayesian bootstrap.

2.1.1. Mid imputation In epidemiological studies including the AIDS studies, individuals are periodically screened for the evidence of infection. It is only known that the exact infection time is 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 for subject i. The simplest method for estimating the infection time is to use the midpoint of the interval (L_i, R_i) (Law and Brookmeyer, 1992).

2.1.2. Conditional mean imputation Conditional mean imputation is an alternative method to reduce the bias of rather simpler imputation method by using the expected infection time given that infection time occurred between L_i and R_i (Gauvreau *et al.*, 1994). Let f be the *pdf* of infection time among the individuals and F be the corresponding *cdf*. Then the imputed infection time of subject i is estimated as

$$\hat{t}_i = \frac{\int_{Li}^{R_i} t\hat{f}(t)dt}{\hat{F}(R_i) - \hat{F}(L_i)}, \quad \text{where } \hat{F}(a) = \int_0^a \hat{f}(t)dt. \tag{2.1}$$

In order to get the imputed infection time we need a parametric model for the infection time.

2.1.3. Approximate Bayesian bootstrap(ABB) Approximate Bayesian bootstrap imputation method was originally proposed by Rubin (1987) to approximate the Bayesian bootstrap for the categorical data. Pan (2000) used ABB for a two-sample test when the data were interval censored. We implement ABB to the HIV infection time which is also interval censored and make the doubly censored incubation time as right censored. We will follow the formulation of Pan (2000) for interval censored data with modification for doubly censored data which we have. The basic scheme of the ABB applied to doubly censored data is given as follows:

Step1: For $m = 1, \ldots, M$,

- Draw two bootstrap samples $\{(U_{ij}^{(m)}, V_{ij}^{(m)})\}$ and $\{(M_{ij}^{(m)}, \delta_{ij}^{(m)})\}$ from each of the interval censored infection time $\{(U_{ij}, V_{ij})\}$ and right censored AIDS onset time $\{(M_{ij}, \delta_{ij})\}$ samples which are given for the two groups. Here, $i = 1, \ldots, n_j$ and $j = 1, 2$, and δ_{ij} is a censoring indicator for the AIDS onset time.
- Estimate the survival functions from the HIV infection time bootstrap samples as $\hat{S}_1^{(m)}$ and $\hat{S}_2^{(m)}$, respectively.
- For each of the two original infection time samples, generate a set of observations $T_{ij}^{(m)}$ as follows. For $j = 1, 2$ and $i = 1, \ldots, n_j$, sample X_{ij} from the distributions $\hat{S}_j^{(m)}$, conditional on that $\{(U_{ij} < X_{ij} \leq V_{ij})\}$ and let $T_{ij}^{(m)} = X_{ij}$.
- Combine the imputed infection time bootstrap samples with the AIDS onset time bootstrap samples.
- Calculate the AIDS incubation time as $Y_{ij}^{(m)} = M_{ij}^{(m)} - T_{ij}^{(m)}$ and obtain a possibly right censored bootstrap samples as $\{(Y_{ij}^{(m)}, \delta_{ij}^{(m)})\}$, where $\delta_{ij}^{(m)}$ is a censoring indicator for $Y_{ij}^{(m)}$.

Step2: Use each $\{(Y_{ij}^{(m)}, \delta_{ij}^{(m)})\}$ to calculate the coefficient estimator $\widehat{\beta^m}$ and its variance estimate $\widehat{\Sigma_*{}^{(m)}}$.

Step3: Let

$$\widehat{\Sigma_*{}^{(m)}} = \frac{1}{M}\widehat{\Sigma_*{}^{(m)}} + \left(1 + \frac{1}{M}\right) \operatorname{var}\left(\widehat{\beta^{(1)}}, \ldots, \widehat{\beta^{(M)}}\right).$$

For estimating the survival functions in Step 1, we used Expectation-Maximization Iterative Convex Minorant(EMICM) estimator of the distribution function proposed by Wellner and Zahn (1997) and used Gibbs sampler to estimate the coefficient.

3. Two Approaches for Estimating AIDS Incubation Period

3.1. Parametric approach

To estimate the covariate effect of the incubation period distribution, we assume the incubation period distribution follows 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*:

$$f(y|a,b) = \left(\frac{a}{b}\right)\left(\frac{y}{b}\right)^{a-1} \exp\left(-\left(\frac{x}{b}\right)^{a}\right), \quad y > 0,\ a > 0,\ b > 0. \tag{3.1}$$

A random variable following Weibull distribution can be denoted as $W(a,b)$, where a is the shape parameter and b is the scale parameter. Formula (3.1) has the same form if we translate $\alpha = a$ and $\gamma = (1/b)^a$ and the *pdf* has the form as (3.2):

$$f(y|\alpha,\gamma) = \alpha\gamma y^{\alpha-1} \exp(-\gamma y^{\alpha}), \quad y > 0,\ \alpha > 0,\ \gamma > 0. \tag{3.2}$$

It is convenient to write (3.2) in terms of the parameterization $\lambda = \log(\gamma)$, leading to

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

To build the Weibull regression model, we introduce the covariates through $\lambda_i = x_i'\beta$. Denote the censoring indicator as $\nu = (\nu_1, \ldots, \nu_n)'$ and the obersved dasta as $D = (n, y, x, \nu)$. Then the joint posterior has the form as (3.3) (Ibrahim *et al.*, 2001):

$$\begin{aligned}\pi(\beta,\alpha|D) \propto \alpha^{\alpha_0+d-1} \exp\Bigg[&\sum_{i=1}^{n} \left\{\nu_i x_i'\beta + \nu_i(\alpha-1)\log(y_i) - y_i^{\alpha}\exp(x_i'\beta)\right\} \\ &- \kappa_0\alpha - \frac{1}{2}(\beta-\mu_0)'\Sigma_0^{-1}(\beta-\mu_0)\Bigg],\end{aligned} \tag{3.3}$$

where α has a gamma prior $G(\alpha_0, \kappa_0)$ and β has a normal prior $N(\mu_0, \Sigma_0)$. For the Weibull regression model the conditional posterior distribution of $[\alpha|\beta, D]$ and $[\beta|\alpha, D]$ are log-concave, and thus the 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 $\{Z(y) : y \geq 0\}$ is called gamma process and is denoted by $Z(y) \sim GP(c\alpha(y), c)$ when it has the properties as follows (Ibrahim *et al.*, 2001):

(i) $Z(0) = 0$

(ii) $Z(y)$ has independent increment disjoint intervals, and

(iii) for $y > s$, $Z(y) - Z(s) \sim G(c(\alpha(y) - \alpha(s)), c)$, where c is the confidence parameter about the mean $\alpha(y)$.

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

$$I_i(y)dy = E\left(dN_i(y)|F_{y^-}\right), \tag{3.4}$$

where $dN_i(y)$ is the increment of N_i over $[y, y + dy)$ and F_{y^-} is the available data just before time y. As $dy \to 0$ (3.4) becomes the instantaneous hazard at time y for subject i. It is assumed to have the proportional hazards form

$$I_i(y) = Y_i(y)\lambda_0(y)\exp(\beta' z_i),$$

where $Y_i(y)$ is an observed process taking value 1 or 0 whether the subject i is observed at time y or not. The joint posterior distribution is defined by

$$P(\beta, \Lambda_0(y)|D) \propto P(D|\beta, \Lambda_0(y))P(\beta)P(\Lambda_0(y)), \tag{3.5}$$

where $D = \{N_i(y), Y_i(y), z_i; i = 1, \dots, n\}$ is the observed data. Looking through the form of the likelihood of $P(D|\beta, \Lambda_0(y))$ in (3.5), $dN_i(y)$ seems as it is independent Poisson random variables with mean $I_i(y)dy$. Hence, for implementing the Gibbs sampler, the variables are assigned prior distributions as follows:

$$dN_i(y) \sim \text{Poisson}(I_i(y)dy), \tag{3.6}$$

where $I_i(y)dy = Y_i(y)\exp(\beta' z_i)d\Lambda_0(y)$. The increment in the integrated baseline hazard function occurring in interval $[y, y + dy)$ can be written as $d\Lambda_0(y) = \lambda_0(y)dy$. The conjugate prior for the Poisson mean is the gamma distribution. Kalbfleisch (1978) suggested a conjugate independent increments prior and it can be written as

$$d\Lambda_0(y) \sim G\left(cd\Lambda_0^*(y), c\right). \tag{3.7}$$

We set $d\Lambda_0^*(y) = rdy$, where r is a prior guess at the failure rate per unit time and dy is the size of the time interval. The parameter c is degree of confidence in prior guess for increment in unknown hazard function.

4. Simulation

We conducted a simulation study to investigate the performance of the imputation methods incorporating prior distributions for the binary covariate effect which indicates two different group. In addition the target range of the simulation study is to investigate the effect of the interval wide effect as well as the censoring amount. The sample size for each group was 50. To generate doubly censored data, we first generated the exact HIV infection time T_i from uniform distribution $U[10, 30]$. The incubation time Y_i for the two group was generated from Weibull distribution with the same shape parameter 2.5 and the scale parameter was set differently with $W(2.5, 10)$ for the first group and $W(2.5, 12)$ for the second group respectively so that $\beta = -0.455$ for the Weibull model and the Cox model. To generate interval censored HIV infection time, the left interval L_i and the right interval R_i was set to T_i plus minus a specified Uniform distribution. In order to investigate the interval width effect we used two Uniform distributions $U[0, 5]$ and $U[0, 7]$. Then the AIDS onset time S_i was computed as $S_i = T_i + Y_i$. The AIDS onset time was set to right censored with two different censoring amounts, moderately censored with 20% and heavily censored with 50% of the data, and we investigated if there was a difference in the performance of the imputation methods based on MC(Monte Carlo) error between the censoring amount. Under four different types of data, we compared the three imputation methods incorporating prior distributions of the covariate effect. The results when the HIV infection time is uncensored are investigated as well. Table 4.1

Table 4.1. Simulation results of uncensored HIV infection time assuming AIDS incubation time follows Weibull distribution with prior $\alpha \sim \exp(0.01)$, $\beta \sim N(0, 1000)$

Mean of β	MC error	2.5%	97.5%
−0.4424	0.0021	−0.8473	−0.0440

Table 4.2. Simulation results of four different data sets with three imputation methods under prior distributions $\alpha \sim \exp(0.01)$, $\beta \sim N(0, 1000)$. AIDS incubation time is assumed to follow Weibull distribution.

Imputation Method	Statistics	Moderate_5	Moderate_7	Heavy_5	Heavy_7
mid point	Mean of β	−0.4335	−0.4102	−0.4562	−0.4225
	MC error	0.0021	0.0020	0.0023	0.0022
	2.5%	0.8382	−0.8090	−0.8609	−0.8229
	97.5%	−0.0328	−0.0111	−0.0500	−0.0217
conditional mean	Mean of β	−0.4380	−0.4167	−0.4594	−0.4209
	MC error	0.0020	0.0021	0.0022	0.0020
	2.5%	−0.8378	−0.8179	−0.8577	−0.8191
	97.5%	−0.0329	−0.0136	−0.0669	−0.0280
ABB	Mean of β	−0.5086	−0.4849	−0.4978	−0.2703
	MC error	0.0026	0.0025	0.0021	0.0020
	2.5%	−0.9192	−0.8930	−0.9009	−0.6678
	97.5%	−0.0928	−0.0770	−0.0972	0.1299

is the result when the infection time is uncensored and the incubation time is assumed to follow a Weibull distribution. For the parametric approach, the prior distribution for parameters of the Weibull model was set to $\alpha \sim \exp(0.01)$, $\beta \sim N(0, 1000)$. Since there is no prior information about the parameters, sufficiently large values of variance for β are allocated to generate noninformative prior distribution. It can be seen from Table 4.1 that the posterior mean of β converges to −0.4424 which is close to the true value of β, −0.466 and has MC error as which is quite small.

For the assessment of the imputation methods incorporating prior distributions of the covariate effect, we investigated the posterior mean of β and its MC error and 95% confidence interval for differently generated data sets. The generated data sets have two different interval width for HIV infection time and two different censoring amounts and thus we have four different data sets. The results are shown in Table 4.2. The four different data sets are noted as Moderate_5, Moderate_7, Heavy_5 and Heavy_7. Moderate_5 is the simulated data when AIDS onset is moderately censored(20%) and two intervals (L_i, R_i) of the HIV infection time is computed as $T_i \pm U[0, 5]$. With the same notation Heavy_7 is the simulated data when AIDS onset is heavily censored(50%) and the two intervals (L_i, R_i) of the HIV infection is computed as $T_i \pm U[0, 7]$. The prior of β is set to $\beta \sim N(0, 1000)$ with a vague precision and α is set to $\alpha \sim \exp(0.01)$. For conditional mean imputation, lognormal distribution $LN(2.82, 0.32)$ was chosen for the HIV infection distribution, and for approximate Bayesian bootstrap(ABB) five data sets were used, that is $M = 5$. It can be seen that the MC error for all the data sets are close to 0.002. Considering the posterior mean, moderately censored data with conditional mean imputation has the closest posterior mean of β to the original value(−0.455) when the interval width is shorter. As the interval width grows larger ABB has the closet posterior mean to the original value. When the data are heavily censored, mid imputation has the posterior mean of β which is closest to the original value for both different interval widths. In general, except for Moderate_7, simple methods like conditional mean imputation and mid imputation performs well. The results in Table 4.2 are similar to those in Table 4.1 except

Table 4.3. Simulation results of uncensored HIV infection time under different prior distributions assuming AIDS incubation time follows Cox's regression with $r = 1$, $c = 0.5$ and $\beta \sim N(0, 1000)$.

Mean of β	MC error	2.5%	97.5%
−0.4677	0.0021	−0.8705	−0.0668

Table 4.4. Simulation results of four different data sets with three imputation methods under prior distributions $r = 1$, $c = 0.5$ and $\beta \sim N(0, 1000)$. AIDS incubation time is assumed to follow Cox's regression model.

Imputation Method	Statistics	Moderate_5	Moderate_7	Heavy_5	Heavy_7
mid point	Mean of β	−0.4365	−0.4314	−0.4999	−0.4572
	MC error	0.0020	0.0020	0.0019	0.0021
	2.5%	−0.8421	−0.8356	−0.9035	−0.8643
	97.5%	−0.0328	−0.0344	−0.0969	−0.0570
conditional mean	Mean of β	−0.4534	−0.4477	−0.4879	−0.4615
	MC error	0.0020	0.0022	0.0020	0.0020
	2.5%	−0.8605	−0.8528	−0.8938	−0.8581
	97.5%	−0.0489	−0.0431	−0.0918	−0.0606
ABB	Mean of β	−0.5168	−0.5237	−0.4423	−0.2879
	MC error	0.0021	0.0022	0.0020	0.0021
	2.5%	−0.9288	−0.9358	−0.8458	−0.6923
	97.5%	−0.1079	−0.1150	−0.0415	0.1146

for ABB with Heavy_7 data, which means that the performance of the imputation methods are satisfactory. We can also see that there is no censoring effect when the data have the same interval width for both mid point imputation and conditional mean imputation. On the other hand as the interval width gets wider, the difference between the posterior mean of β and the original value gets larger when the data have the same amount of censoring for both mid imputation and conditional mean imputation. As for the ABB, there seems to be no interval width effect but it has censoring effect.

We next assume that the AIDS incubation time follows Cox's regression model. For prior distribution, β was also set to $\beta \sim N(0, 1000)$ which is the same as the parametric approach for noninformative prior distribution. For other prior distributions, we must decide r and c, defined in Section 3.2, which is a prior guess at failure rate and degree of confidence in prior guess for increment in unknown hazard function, respectively. We set $r = 1$ and $c = 0.5$ considering a situation with weak confidence for prior guess of increment. The results are shown in Table 4.3 when the HIV infection time is uncensored. In Table 4.3, we can see that the posterior mean converges to −0.4677 which is close to the true value −0.455 with MC error 0.0021.

The results for different imputation methods are shown in Table 4.4. First we can see the MC error for all the data sets are close to 0.002, which gives the same result as when the AIDS incubation period is assumed to follow Weibull distribution.

Considering the posterior mean, when the data is moderately censored, conditional mean imputation has the closest posterior mean value of β for both interval widths. On the other hand for the heavily censored data, ABB and mid imputation has the closest posterior mean value of β for shorter and wider HIV interval width respectively. Investigating the interval width effect and censoring effect, for conditional mean imputation there are no either effects but for mid imputation and for ABB there seems to be an interaction effect between the two effects. For all the simulation data, with both parametric and semi-parametric approach, in order to investigate the convergence of the posterior

Table 5.1. Estimators for each imputation method with parametric approach for AIDS Cohort Study under prior distributions $\alpha \sim \exp(0.1)$, $\beta \sim N(0.7, 0.0001)$

Imputation Method	Mean of $\hat{\beta}$	MC error	2.5%	97.5%
mid	0.6760	$5.95E-5$	0.6567	0.6958
conditional mean	0.6759	$6.13E-5$	0.6564	0.6954
ABB	0.6759	$5.94E-5$	0.6566	0.6954

mean we set 3 chains with different initial values and updated the model for 10000 iterations and monitored the convergence of the MCMC simulation. The figures of multiple chain trace plots and Gelman_Rubin convergence statistics are not shown here, but for all the imputation methods, the multiple chains mix rapidly and the potential scale reduction estimator was near 1.

5. Real Example

We also applied our methods to the real data discussed in Kim *et al.* (1993) from the AIDS cohort study of hemophiliacs, ane investigated the performance of the imputation methods under the different priors of the covariate effect. This study consists of the individuals with Type A or B hemophilia who were at risk of HIV infection through the contaminated blood factor they received. The subjects were classified into two groups according to the amount of blood they received. The data have 188 HIV infected subjects with all interval censored infection time and right censored AIDS onset time. The censoring rate of the incubation time is 82% and the length of the HIV infection width is mostly under 10.

For the parametric approach we assume that the incubation time follows the Weibull distribution. We introduce the covariate through $\lambda_i = \beta x_i$ and the parameter we are interested in is β, which is the coefficient of the treatment level. Before estimating β we imputed the interval censored HIV infection time using mid imputation, conditional mean imputation and ABB multiple imputation. After imputing we could obtain the right censored AIDS incubation time. To estimate the covariate effect using Gibbs sampler, we first setup β to follow $\beta \sim N(0.7, 0.001)$. The prior mean for β is chosen with strong confidence since the variance is near zero. This prior information about β was achieved by other literatures (Kim *et al.*, 1993; Sun *et al.*, 1999) which analyzed the same AIDS cohorts data using Cox regression. The methods they have used to estimate the covariate effect of the level of the treated group are different but the estimates are all close to 0.7 (0.69 obtained by Kim *et al.*, 1993 and 0.7039 obtained by Sun *et al.*, 1999). We used this prior for β in both parametric approach and semi-parametric approach. For the parametric approach assuming that the AIDS incubation period follows Weibull distribution, the prior distribution for α was given $\alpha \sim \exp(0.1)$. The results are given in Table 5.1.

As shown in Table 5.2, all three imputation methods have posterior mean of β as 0.676. This value is close to those obtained by Kim *et al.* (1993) and Sun *et al.* (1999). When we compare the three imputation methods, ABB has the smallest MC error ($5.94E-5$) and mid imputation has quite the same MC error ($5.95E-5$) as ABB. On the other hand conditional mean imputation has the largest MC error which is ($6.13E-5$) among the three imputation methods but the difference among them are quite small. The posterior mean for β is 0.6760 for all imputation methods indicating that the heavily treated group has shorter incubation period than the lightly treated group.

For the semi-parametric approach, gamma process was used and we fixed the prior guess of failure

Table 5.2. Estimators for each imputation method with semi-parametric approach for AIDS Cohort Study under prior distributions $r = 0.1$, $c = 0.01$, $\beta \sim N(0.7, 0.0001)$

Imputation Method	Mean of $\hat{\beta}$	MC error	2.5%	97.5%
mid	0.7204	$1.85E-4$	0.6585	0.7821
conditional mean	0.7189	$2.03E-4$	0.6575	0.7791
ABB	0.7000	$6.23E-5$	0.6804	0.7197

rate as $\gamma = 0.01$ and the degree of confidence in prior guess for increment in unknown hazard function parameter as $c = 0.1$. As shown in Table 5.2, ABB has the smallest MC error ($6.23E-5$) and the conditional mean imputation has the largest MC error ($1.85E-4$). The posterior mean of β for all the three imputations are near 0.7 which also indicate that the heavily treated group had shorter incubation period than the lightly treated group. For both results assuming Weibull distribution and Cox's regression model, the posterior means of β are quite similar between groups, but when considering the MC error ABB has the smallest value and the conditional mean imputation has the largest value.

For the real example, we also set 3 multiple chains with different initial values and monitored the convergence of the MCMC simulation. For all the imputation methods the multiple chains mix rapidly and also the potential scale reduction estimator was near 1.

6. Conclusion and Discussion

This paper considered estimation of the covariate effect when the data are doubly censored. It is difficult to analyze the doubly censored data, and thus there have been many imputing methods proposed. Imputing makes the original data easy to analyze with the existing statistical methods. We compared three imputation methods: Mid imputation and conditional mean imputation for single imputation methods and ABB for multiple imputation method. After imputing the interval censored HIV infection time, two approaches for estimating the AIDS incubation time were applied. Weibull model was applied to the parametric approach and gamma process was applied to the semi-parametric approach. After imputing the incubation time we proposed to estimate the covariate effect of AIDS incubation period through Bayesian approach using Gibbs sampler. We compared the results of the 3 imputation methods under certain prior distribution of the parameters. In order to investigate the effect of the HIV infection time interval width and the severity of the censoring amount of AIDS onset period, a simulation study was done. The simulation study shows that, both for the parametric approach and the semi-parametric approach, rather simpler imputation methods such as mid imputation and conditional mean imputation performs well rather than ABB. Also the interval width effect and the censoring effect depends on the assumption of the AIDS incubation period. As for the real example ABB had the best performance respect to the MC error in both parametric and semi-parametric approach between the three imputation methods. This result is the same when the simulation setting is Heavy_5 since in the real data example, the data is also heavily censored and the length of the HIV infection width is mostly of 10. The generalization of our results can be of limited since the results can depend on the data structure and the effect of the truncation. However from our simulation study and through the real example, simple imputation methods like mid imputation and conditional mean imputation also perform as well as quite complex imputation methods like ABB. We have also shown how to incorporate the prior distribution in estimating the covariate effect when AIDS incubation period are doubly censored. With using a Bayesian approach, flexible modeling and the use of prior information is available. Further works might be to compare through Bayesian paradigm with other imputation methods in nonparametric survival models with mixing models as well.

References

Arjas, E. and Gasbarra, D. (1994). Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler, *Statistica Sinica*, **4**, 505–524.

Berliner, L. M. and Hill, B. M. (1988). Bayesian nonparametric survival analysis, *Journal of the American Statistical Association*, **83**, 772–779.

Brookmeyer, R. and Goedert, J. (1989). Censoring in an epidemic with an application to hemophilia-associated AIDS, *Biometrics*, **45**, 325–335.

Burridge, J. (1981). Empirical Bayes analysis of survival time data, *Journal of the Royal Statistical Society, Series B*, **43**, 65–75.

De Gruttola, V. G. and Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS, *Biometrics*, **45**, 1–11.

Gauvreau, K., DeGruttola, V., Pagano, M. and Bellocco, R. (1994). Markers and incubation time: The effect of covariates on the induction time of AIDS using improved imputation of exact seroconversion times, *Statistics in Medicine*, **13**, 2021–2030.

Geskus, R. B. (2001). Methods for estimating the AIDS incubation time distribution when date of seroconversion is censored, *Statistics in Medicine*, **20**, 795–812.

Gomez, G. M. and Calle, M. L. (1999). Nonparametric estimation with doubly censored data, *Journal of Applied Statistics*, **26**, 45–58.

Ibrahim, J. G., Chen, M. H. and Sinha, D. (2001). *Bayesian Survival Analysis*, Springer-Verlag, America.

Kalbfleisch, J. D. (1978). Nonparametric Bayesian analysis of survival time data, *Journal of the Royal Statistical Society, Series B*, **40**, 214–221.

Kim, M. Y., De Gruttola, V. G. and Lagakos, S. W. (1993). Analyzing doubly censored data with covariates, with application to AIDS, *Biometrics*, **49**, 13–22.

Law, C. G. and Brookmeyer, R. (1992). Effects of mid-point imputation on the analysis of doubly censored data, *Statistics in Medicine*, **11**, 1569–1578.

Pan, W. (2000). A two-sample test with interval censored data via multiple imputation, *Statistics in medicine*, **19**, 1–11.

Pan, W. (2001). A multiple imputation approach to regression analysis for doubly censored data with application to AIDS studies, *Biometrics*, **57**, 1245–1250.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

Sinha, D. and Dey, D. (1997). Semiparametric Bayesian analysis of survival data, *Journal of the American Statistical Association*, **92**, Review paper.

Sun, J. (1995). Empirical estimation of a distribution function with truncated and doubly interval-censored data and its application to AIDS studies, *Biometrics*, **51**, 1096–1104.

Sun, J. (2004). Statistical analysis of doubly interval-censored failure time data, In Handbook of Statistics 23: Advances in Survival Analysis (Eds., N. Balakrishnan and C. R. Rao), 105–122, North-Holland, Amsterdam, The Netherlands.

Sun, J., Liao, Q. and Pagano, M. (1999). Regression analysis of doubly censored failure time data with applications to AIDS studies, *Biometrics*, **55**, 909–914.

Wellner, J. A. and Zhan, Y. (1997). A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data, *Journal of the American Statistical Association*, **92**, 945–959.