HHS Public Access

Author manuscript

Comput Stat Data Anal. Author manuscript; available in PMC 2016 November 01.

Published in final edited form as:

Comput Stat Data Anal. 2016 November ; 103: 242–249. doi:10.1016/j.csda.2016.05.011.

A multiple imputation approach to the analysis of clustered interval-censored failure time data with the additive hazards model

Ling Chen[a,*], **Jianguo Sun**[b], and **Chengjie Xiong**[a]

[a]Division of Biostatistics, Washington University School of Medicine, Campus Box 8067, 660 S. Euclid Ave, St. Louis, MO 63110, United States

[b]Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211, United States

Abstract

Clustered interval-censored failure time data can occur when the failure time of interest is collected from several clusters and known only within certain time intervals. Regression analysis of clustered interval-censored failure time data is discussed assuming that the data arise from the semiparametric additive hazards model. A multiple imputation approach is proposed for inference. A major advantage of the approach is its simplicity because it avoids estimating the correlation within clusters by implementing a resampling-based method. The presented approach can be easily implemented by using the existing software packages for right-censored failure time data. Extensive simulation studies are conducted, indicating that the proposed imputation approach performs well for practical situations. The proposed approach also performs well compared to the existing methods and can be more conveniently applied to various types of data representation. The proposed methodology is further demonstrated by applying it to a lymphatic filariasis study.

Keywords

Additive hazards model; Clustered interval-censored data; Multiple imputation; Within-cluster resampling

1. Introduction

Correlated or clustered failure time data occur in many areas such as medical studies (Cai and Prentice, 1997). In many situations, the exact time of failure cannot be observed but is known to fall between two observation times. Such data are usually referred to as interval-censored failure time data. Clustered interval-censored failure time data may naturally arise in periodic follow-up studies where each study subject is repeatedly measured at discrete time points and some subjects come from the same cluster or group such as sibling, family

*Corresponding author. Tel.: +1 314 747 2373; fax: +1 314 362 2693. lingchen@wustl.edu (L. Chen) . sunj@missouri.edu (J. Sun), chengjie@wustl.edu (C. Xiong).

and community. In this case, the interval-censored observations from the same cluster share certain unknown characteristics and are correlated as a result.

Several methods have been proposed for regression analysis of clustered interval-censored failure time data. These include parametric approaches (Bellamy et al., 2004; Zhang and Sun, 2010; Lam et al., 2010) and semiparametric approaches (Xiang et al., 2011; Zhang and Sun, 2013; Li et al., 2014). For most of the existing semiparametric methods, it is assumed that the failure time of interest follows the proportional hazards model (Pan, 2000). It is known that the proportional hazards model may not fit failure time data well sometimes and in this case, one of the alternatives is the additive hazards model. The additive hazards model describes a different aspect of the relationship between survival time and covariates and could be more plausible than the proportional hazards model in many situations (Lin and Ying, 1994). For example, in public health studies, people may be more interested in the risk difference described by the additive hazards model than the risk ratio described by the proportional hazards model (Breslow and Day, 1987). Li et al. (2012) investigated regression analysis of clustered interval-censored failure time data. However, they considered only the simple situation where each subject is observed only twice, or the failure time is left-, interval-, or right-censored. In the following, we will discuss general clustered interval-censored failure time data.

We develop a multiple imputation approach to analyze clustered interval-censored failure time data. Our method reduces the analysis of clustered interval-censored data to that of right-censored failure time data. The algorithm is implemented in two steps. First, we apply a within-cluster resampling (WCR) procedure (Hoffman et al., 2001) to sample a single subject from each cluster and obtain independent interval-censored failure time data. In other words, one observation is randomly sampled with replacement from each of the N clusters and an independent interval-censored dataset of size N is formed. Then a multiple imputation procedure is applied to convert the interval-censored dataset to right-censored failure time data. Inference is made based on the right-censored failure time data taking advantage of an existing estimation procedure developed by Lin and Ying (1994). If the resampling process is repeated a large number of times, say Q, where each of the Q analyses provides a consistent estimator of parameter of interest, one can obtain the WCR estimate by taking the average of the Q resampling-based estimates.

The multiple imputation approach has been used in many fields. In particular, Lam et al. (2010) developed such a procedure for the same type of the data considered here but under a parametric gamma frailty model. In addition, their method requires some strong assumptions on the distribution of the unobserved frailty and the use of the EM algorithm for the implementation. In contrast, the presented procedure has the advantage of simplicity because it can be easily implemented by using a standard software for the additive hazards model. Another advantage of the proposed approach is the use of a semiparametric additive hazards model, which provides more flexibility in describing the relationship between the failure time and covariates than a fully parametric model. In addition, the procedure given below does not need the estimation of the unobserved frailty.

In the following, before presenting the multiple imputation approach, we will first briefly introduce in Section 2 the additive hazards model and the inference procedure proposed by Lin and Ying (1994) for independent right-censored failure time data. The proposed estimation procedure is then presented in Section 3 and as mentioned above, it makes use of the method given in Lin and Ying (1994). Results from an extensive simulation study are reported in Section 4 for assessing the performance of our proposed approach and comparing to the existing method. Section 5 applies the proposed method to a set of well-known clustered interval-censored failure time data arising from a lymphatic filariasis study. Section 6 contains some concluding remarks.

2. The additive hazards model and right-censored failure time data with independent samples

Consider a survival study and let T denote the failure time of interest and Z a vector of covariates that may depend on time t. We assume that given Z, the hazard function of T has the form

$$\lambda(t;Z) = \lambda_0(t) + \beta' Z(t), \quad (1)$$

where $\lambda_0(t)$ denotes the unknown baseline function and β is the vector of unknown regression coefficients. That is, T follows the additive hazards model (Cox and Oakes, 1984). In the following, as most authors, it will be assumed that the covariate $Z(t)$ is known or can be observed at any time.

In this section, we will assume that instead of clustered data, one observes right-censored failure time data given by $\{X_i, \delta_i, Z_i, i = 1, \ldots, n\}$ from n independent subjects. Here X_i denotes the observed failure time defined as the minimum of the true failure time T_i and the censoring time for subject i and $\delta_i = 1$ if the true failure time is observed and 0 otherwise. Also it will be assumed that the failure time and the censoring time are independent given covariates. Define $Y_i(t) = I(X_i \geq t)$, the risk indicator process, and $N_i(t) = I(X_i \leq t, \delta_i = 1)$, a counting process, $i = 1, \ldots, n$.

To estimate β in model (1), Lin and Ying (1994) proposed to use the following estimating equation

$$U(\beta) = \sum_{i=1}^{n} \int_0^\infty \left\{ Z_i(t) - \bar{Z}(t) \right\} \left\{ dN_i(t) - Y_i(t) \beta' Z_i(t)\, dt \right\} = 0,$$

where

$$\bar{Z}(t) = \frac{\sum_{i=1}^{n} Y_i(t) Z_i(t)}{\sum_{i=1}^{n} Y_i(t)}.$$

It can be easily shown that the solution to the equation above has the explicit form

$$\hat{\beta}=\left[\sum_{i=1}^{n}\int_0^{\infty}Y_i(t)\left\{Z_i(t)-\bar{Z}(t)\right\}^{\otimes 2}dt\right]^{-1}\left[\sum_{i=1}^{n}\int_0^{\infty}\left\{Z_i(t)-\bar{Z}(t)\right\}dN_i(t)\right], \quad (2)$$

where $a^{\otimes 2}=a\quad a'$ for a vector a. Furthermore, Lin and Ying (1994) showed that $n^{1/2}\left(\hat{\beta}-\beta_0\right)$ converges weakly to a normal vector with mean zero and a covariance matrix that can be consistently estimated by $\Sigma = A^{-1}\,B\,A^{-1}$, where β_0 denotes the true value of β,

$$A=\frac{1}{n}\quad\sum_{i=1}^{n}\int_0^{\infty}Y_i(t)\left\{Z_i(t)-\bar{Z}(t)\right\}^{\otimes 2}dt,$$

and

$$B=\frac{1}{n}\quad\sum_{i=1}^{n}\int_0^{\infty}\left\{Z_i(t)-\bar{Z}(t)\right\}^{\otimes 2}dN_i(t).$$

Given $\hat{\beta}$, a natural estimate of the baseline cumulative hazard function $\Lambda_0(t)=\int_0^t\lambda_0(s)\,ds$ is given by

$$\hat{\Lambda}_0\left(t;\hat{\beta}\right)=\int_0^t\frac{\sum_{i=1}^{n}\left\{dN_i(s)-Y_i(s)\,\hat{\beta}'\,Z_i(s)\,ds\right\}}{\sum_{i=1}^{n}Y_i(s)}. \quad (3)$$

Note that the estimator (3) may not always be monotone in t. For this, Lin and Ying (1994) suggested using $\hat{\Lambda}_0^{*}\left(t;\hat{\beta}\right)=max_{s\le t}\quad\hat{\Lambda}_0\left(s;\hat{\beta}\right)$. In the next section, we will discuss the use of the estimators (2) and (3) in estimation of β when one observes clustered interval-censored failure time data.

3. A multiple imputation approach for estimation of β with clustered interval-censored failure time data

Now we consider a survival study that includes n independent clusters with n_i denoting the size of the ith cluster, where $i = 1, \ldots, n$. Let T_{ij} denote the failure time of interest for subject j in cluster i, $j = 1, \ldots, n_i$. Suppose that one observes clustered interval-censored failure time data on the T_{ij}'s denoted by $\{(L_{ij}, R_{ij}], Z_{ij}(t), i = 1, \ldots, n; j = 1, \ldots, n_i\}$, where $L_{ij} \le R_{ij}$. Here $L_{ij} = 0$ represents a left-censored observation, $R_{ij} = \infty$ corresponds to a right-censored one, and $Z_{ij}(t)$ is a possibly time-dependent covariate vector associated with subject j in cluster i. Note that for interval-censored data, all available information about T_{ij} is the interval $(L_{ij}, R_{ij}]$ with $T_{ij} \in (L_{ij}, R_{ij}]$. We assume that L_{ij} and R_{ij} are independent of T_{ij} given

the covariates $Z_{ij}(t)$. Throughout this paper the $T_{ij}^{'}$s are considered to be independent for subjects in different clusters but could be dependent for subjects within the same cluster.

To describe the association within each cluster, we will assume that there exists a cluster-specific latent variable b_i with mean zero and unknown variance. Furthermore suppose that given covariates and b_i, the failure times within each cluster are independent and T_{ij} follows the additive hazards model

$$\lambda_{ij}(t \mid Z_{ij}, b_i) = \lambda_0(t) + \beta' Z_{ij}(t) + b_i. \quad (4)$$

Here $\lambda_0(t)$ and β are defined as in model (1). To make inference about β, we propose to randomly sample with replacement one interval-censored observation from each cluster using the WCR approach (Hoffman et al., 2001) and construct a new independent sample of interval-censored observations. The left-censored and interval-censored observations in this resampled interval-censored data will be imputed to generate a right-censored failure time data. Then the problem is reduced to analyzing the imputed (right-censored) data, which can be handled with the inference procedure discussed in Section 2.

More specifically, we randomly select one data point $\{(L_i, R_i], Z_i(t)\}$ with replacement from each cluster and generate a set of interval-censored failure time data. Let Q be a positive integer denoting the number of resamples. We repeat the resampling process Q times and generate Q sets of independent samples of interval-censored failure time data. For each set of the interval-censored failure time data, we propose the following estimation procedure to estimate β. Let K be a prespecified integer representing number of imputations.

Step 1. Choose initial estimates $\hat{\beta}^{(0)}$ and $\hat{S}_0^{(0)}$ of β and the baseline survival function $S_0(t) = \exp\{-\Lambda_0(t)\}$, respectively.

Step 2. At the lth iteration, let $\hat{\beta}^{(l-1)}$ and $\hat{S}_0^{(l-1)}$ denote the estimates of β and $S_0(t)$ obtained at the $(l-1)$th iteration and generate K sets of right-censored data $\{X_{ik}, \delta_{ik}, Z_{ik}, i = 1, \ldots, n; k = 1, \ldots, K\}$ as follows. If $R_i = \infty$ (right-censored), define $X_{ik} = L_i$ and $\delta_{ik} = 0$; if $R_i < \infty$, define $\delta_{ik} = 1$ and X_{ik} to be a random number drawn from the survival function

$$\hat{S}_0^{(l-1)}\left(t; \hat{\beta}^{(l-1)}\right) \quad exp\left\{-\hat{\beta}^{(l-1)'} Z_i^*(t)\right\} \quad (*)$$

given $L_i \le X_{ik} < R_i$, where $Z_i^*(t) = \int_0^t Z_i(s)\, ds$. The imputed value was drawn this way: suppose that, in interval $(L_i, R_i]$, the survival function (*) has probability mass $\{p_1, \ldots, p_{S_i}\}$ at time points $\{t_1, \ldots, t_{S_i}\}$, then X_{ik} was randomly drawn from $\{t_1, \ldots, t_{S_i}\}$ with probability proportional to $\{p_1, \ldots, p_{S_i}\}$. Also for all i and k, define $Z_{ik} = Z_i$.

Step 3. First define the estimate $\hat{\beta}_k^{(l)}$ as $\hat{\beta}$ given in (2) with $\{X_i, \delta_i, Z_i, i = 1, \ldots, n\}$ replaced by $\{X_{ik}, \delta_{ik}, Z_{ik}, i = 1, \ldots, n\}$. Then determine the estimate

$\hat{S}_{0k}^{(l)}\left(t;\hat{\beta}_k^{(l)}\right)=exp\left\{-\hat{\Lambda}_{0k}^{(l)}\left(t;\hat{\beta}_k^{(l)}\right)\right\}$, where $\hat{\Lambda}_{0k}^{(l)}\left(t;\hat{\beta}_k^{(l)}\right)$ is given by (3) with $\{X_i, \delta_i, Z_i, i = 1, \ldots, n\}$ replaced by $\{X_{ik}, \delta_{ik}, Z_{ik}, i = 1, \ldots, n\}$. Also calculate the covariance matrix $\Sigma_k^{(l)}$ as Σ given in the previous section based on $\{X_{ik}, \delta_{ik}, Z_{ik}, i = 1, \ldots, n\}$.

Step 4. Define the updated regression estimate $\hat{\beta}^{(l)}$ and the estimate of baseline survival $\hat{S}_0^{(l)}$ as

$$\hat{\beta}^{(l)}=\frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_k^{(l)}, \quad \hat{S}_0^{(l)}\left(t;\hat{\beta}^{(l)}\right)=\frac{1}{K}\sum_{k=1}^{K}\hat{S}_{0k}^{(l)}\left(t;\hat{\beta}_k^{(l)}\right),$$

and the covariance matrix can be estimated by

$$\hat{\Sigma}^{(l)}=\frac{1}{K}\sum_{k=1}^{K}\hat{\Sigma}_k^{(l)}+\left(1+\frac{1}{K}\right)\sum_{k=1}^{K}\frac{\left(\hat{\beta}_k^{(l)}-\hat{\beta}^{(l)}\right)\left(\hat{\beta}_k^{(l)}-\hat{\beta}^{(l)}\right)'}{K-1}.$$

Step 5. Return to step 2 until the convergence is achieved.

To implement the procedure above, one needs to choose the initial estimates of β and S_0. For β, a simple choice is to let $\hat{\beta}^{(0)}=0$ or use the estimate (2) based on the right-censored data $\{X_i, \delta_i, Z_i, i = 1, \ldots, n\}$. To generate the X_i's, one can simply impute $X_i = (L_i + R_i)/2$ for the left- and interval-censored observations. Another alternative is to generate a uniform random variable from the interval $(L_i, R_i]$. For $\hat{S}_0^{(0)}$, as with the second approach for β, one could apply the estimate (3) based on the same set of right-censored data and the relationship $S_0(t) = \exp\{-\Lambda_0(t)\}$. For the convergence of the procedure, since the main purpose is to estimate β, one can apply the criterion

$$\left|\frac{\hat{\beta}^{(l)}-\hat{\beta}^{(l-1)}}{\hat{\beta}^{(l-1)}}\right|\leq\epsilon \quad \text{or} \quad \left|\hat{\beta}^{(l)}-\hat{\beta}^{(l-1)}\right|\leq\epsilon$$

for a given positive constant ϵ.

Note that as the estimator (3), the estimator $\hat{S}_0^{(l)}\left(t;\hat{\beta}^{(l)}\right)$ may not be monotone. In this case, we replace it by $\min_{s\leq t}\hat{S}_0^{(l)}\left(s;\hat{\beta}^{(l)}\right)$ in *Step* 4. Let $\hat{\beta}=\hat{\beta}^{(L)}$ denote the estimate of β at convergence with the multiple imputation procedure above for a resampled interval-censored data. When n is large, the distribution of $\hat{\beta}$ can be approximated by the normal distribution (Rubin, 1987) with mean β_0 and the covariance matrix that can be estimated by

$$\hat{\Sigma}^{(L)}=\frac{1}{K}\sum_{k=1}^{K}\hat{\Sigma}_k^{(L)}+\left(1+\frac{1}{K}\right)\sum_{k=1}^{K}\frac{\left(\hat{\beta}_k^{(L)}-\hat{\beta}^{(L)}\right)\left(\hat{\beta}_k^{(L)}-\hat{\beta}^{(L)}\right)'}{K-1}.$$

Let $\hat{\beta}_q^{(L)}$ and $\hat{\Sigma}_q^{(L)}$ denote the final estimate of β and Σ, respectively, for the qth resampled interval-censored failure time data with $q = 1, \ldots, Q$. Following Hoffman et al. (2001), we estimate β by

$$\hat{\beta}=\frac{1}{Q}\sum_{q=1}^{Q}\hat{\beta}_q^{(L)}$$

and Σ by

$$\hat{\Sigma}=\frac{1}{Q}\sum_{q=1}^{Q}\hat{\Sigma}_q^{(L)}-\frac{1}{Q}\sum_{q=1}^{Q}\left(\hat{\beta}_q^{(L)}-\hat{\beta}\right)\left(\hat{\beta}_q^{(L)}-\hat{\beta}\right)'.$$

In the numerical studies reported below, we used Fortran 77 for the implementation of the computational algorithm above on Redhat (2.6.9–34.EL), and the computational cost seems to be affordable for moderate sample sizes. For example, with a dataset of 100 clusters with the cluster size being between 2 and 5, it took about 50 s of CPU time on average on Dell Desktop (OPTIPLEX 7010 of 10 GB RAM, 1 TB hard disk) to achieve convergence with 10 imputations and 1000 resamplings.

4. A simulation study

An extensive simulation study is conducted to assess the performance of the multiple imputation approach presented in the previous sections. In the study, we investigated the performance of the imputation method with clustered interval-censored observations represented in $(L_{ij}, R_{ij}]$, $i = 1, \ldots, n, j = 1, \ldots, n_i$. For comparison, we also studied the estimating equation-based procedure proposed by Li et al. (2012) under the same additive hazards model. For the generation of the censoring intervals, we considered both the situation where they are independent of covariates and the situation where they may be covariate-dependent.

To generate the failure time of interest, we took $\lambda_0(t) = 2$ in model (4) and generated the Z_{ij}'s from the Bernoulli distribution with success probability $p = 0.5$. The latent variable b_i was assumed to follow a normal distribution with mean zero and variance $\frac{1}{4}$. For the generation of covariate-independent censoring intervals, we first generated multiple examination times for each subject to mimic many medical follow-up studies. For example, this would be the case if each subject is repeatedly monitored at discrete time points and some subjects are siblings or from the same family. More specifically, it was supposed that there were potentially eight examination times for each subject and the length or period

between two consecutive examination times, denoted as *len*, was assumed to be a uniform random variable over (0, 0.2). Denote $\tau_0 = 0$ and $\tau_9 = +\infty$ for each subject, and suppose that τ_{1ij} is the first random examination time generated from the uniform distribution $U(0, 0.2)$. Then the following seven examination times were calculated as $\tau_{kij} = \tau_{1ij} + (k-1) \times len$, $k = 2, 3, \ldots, 8$. It was assumed that a subject may miss the scheduled examination times with probability 0.1 for the first four and 0.2 for the latter four. Then the observed interval $(L_{ij}, R_{ij}]$ was defined with $L_{ij} = \tau_{sij}$ and $R_{ij} = \tau_{qij}$, where τ_{sij} and τ_{qij} with $0 \le s < q \le 9$ are the two real adjacent observation times such that (τ_{sij}, τ_{qij}) contains the true failure time T_{ij}.

For the generation of covariate-dependent censoring intervals, we used the same procedure as above except that the eight examination times τ_k, $k = 1, \ldots, 8$ were assumed to follow the Cox models

$$\lambda_{ij}^{k}(t \mid Z_{ij}(s), s \le t) = \lambda_k(t) \exp\left\{\gamma' Z_{ij}(t)\right\}$$

with $\lambda_k(t) = 9 - k$ and under the restricted order $\tau_{1ij} < \tau_{2ij} < \tau_{3ij} < \tau_{4ij} < \tau_{5ij} < \tau_{6ij} < \tau_{7ij} < \tau_{8ij}$. Here as before, we also assume $\tau_{0ij} = 0$ and $\tau_{9ij} = +\infty$. For all situations, the cluster size n_i was assumed to follow a uniform distribution $U\{2, 3, 4, 5\}$.

In the simulation study, for comparison, we also considered the estimating equation-based procedure given by Li et al. (2012), which assumed that each subject is observed twice at U_{ij} and V_{ij} with $U_{ij} \le V_{ij}$ and T_{ij} is known only $T_{ij} \le U_{ij}$, $U_{ij} < T_{ij} \le V_{ij}$ or $T_{ij} > V_{ij}$. That is, the observation on T_{ij} is left-, interval- or right-censored. It is apparent that such data cannot give exact or right-censored data as the type of interval-censored considered here. To transform the observations with format $(L_{ij}, R_{ij}]$ to that with (U_{ij}, V_{ij}), we set $U_{ij} = R_{ij}$ and let V_{ij} to be the largest observation times in the study if L_{ij} 0. In the case of $0 < L_{ij} < R_{ij} < +\infty$, we set $U_{ij} = R_{ij}$ and let V_{ij} to be the largest observation times in the study if $L_{ij} = 0$. In the case of $0 < L_{ij} < R_{ij} < +\infty$, we set $U_{ij} = L_{ij}$ and $V_{ij} = R_{ij}$ and if $R_{ij} = +\infty$, we took $V_{ij} = L_{ij}$ and U_{ij} to be the smallest observation time in the study. The results given below are based on 1000 replications with $K = 10$ and $Q = 1000$ for the multiple imputation approach.

Table 1 presents the simulation results for estimation of β with the true value, denoted by $\beta 0$, of β taken to be −0.25, 0 and 0.25 with covariate-independent censoring intervals. The results with covariate-dependent censoring intervals are given in Tables 2 and 3 with the true value $\gamma_0 = -0.25$ and 0.25, respectively. In both cases, we considered $n = 100$ and 200, and the results include the bias of β (BIAS), the means of the estimated standard deviations (ESD), the sample standard errors of the estimated β (SSE), and the 95% empirical converge probabilities (CP). It can be seen from the three tables that the imputation approach performs well as the estimator has negligible bias and the estimated standard deviation is close to the empirical standard error with proper coverage probability. But the estimates given in Li et al. (2012) seem to be biased and the biases appear to be more severe with covariate-dependent censoring intervals. A possible reason for this is that the data structure may not satisfy the requirement needed in Li et al. (2012). One can also see from the tables that as expected, both bias and variance decrease when the number of cluster increases.

Note that in the above, we only considered the situation with one covariate. As suggested by a reviewer, we also performed some simulations with two covariates and Table 4 presents some results on estimation of two regression parameters β_1 and β_2 with Z_1 and Z_2 generated from the Bernoulli distribution with success probability $p = 0.5$ and the uniform distribution U(0, 5), respectively. Here the censoring intervals were assumed to be covariate-independent and all other set-ups were the same as with Table 1. The results again indicate that the presented approach seems to work well for the situations considered. We also investigated other set-ups and obtained similar results. Also suggested by a reviewer, we also investigated some different ways for generating covariate-dependent censoring intervals. In the results given in Table 5, for example, the observation times and censored intervals were generated as in Table 1, but with the length *len* following the uniform distribution $U(0, 0.2)$ if $Z \leq 0.5$ and the uniform distribution $U(0.2, 0.8)$ if $Z > 0.5$. Note that here the covariate Z was assumed to follow the uniform distribution $U(0, 1)$. One can see that the results are similar to those given in Tables 2 and 3.

For the simulation results given above, we used $Q = 1000$ for resampling. To investigate the effect of Q on the performance of the procedure, we used $Q = 2000$ and obtained similar results. As mentioned above, in practice, one can approximate the distribution of the proposed estimator $\hat{\beta}$ by the normal distribution. To see this, based on the simulated data generated above, we obtained and evaluated the quantile plots of the standardized $\hat{\beta}$ against the standard normal variable. Figs. 1 and 2 present two such plots from the set-ups in Tables 1 and 3, respectively. Both plots suggest that the normal approximation seems to be reasonable and the plots for other set-ups were similar.

5. An application

In this section, we apply the presented estimation procedure to a set of clustered interval-censored failure time data arising from a lymphatic filariasis (LF) study (Dreyer et al., 2006). It followed 47 men with LF, a debilitating parasitic disease that worms live in nests in human body. Among them, 22 received the co-administration of diethylcarbamazine (DEC) and albendazole (ALB) (new treatment), and the others were given DEC alone (standard treatment). ALB is an anti-parasitic drug, which is commonly used to treat interstim worm infections. When co-administered with DEC, it helps break the cycle of LF transmission between mosquitoes and humans. Using ultrasound a doctor can detect the movement of the living adult worms. One main goal of the study was to compare the effect of the co-administration of DEC and ALB versus DEC alone for the treatment of LF. The variable of interest is the time to clearance of worms in each nest. The patients in the study were followed for one year since their treatment and periodically examined by ultrasound to see if the worms were still alive. Therefore only interval-censored data were observed with each patient as a cluster with $n = 47$ and the cluster size is number of nests of adult filial worms in each patient. In total, 78 adult worm nests were detected by ultrasound and the cluster size, n_i, ranged from 1 to 5. Two covariates were of particular interest to the investigators. One is the treatment indicator, and the other is age of the patient in years at baseline. In the analysis, we define X_{1i} to be 0 if the patient i was given the co-administration of DEC and ALB and 1 otherwise and let X_{2i} be age of the patient at baseline, $i = 1, \ldots, 47$. Note that

here we only have cluster-specific covariates. The data are given in the formulation $T_{ij} \in (L_{ij}, R_{ij}]$.

To analyze the data, we applied the multiple imputation procedure with $Q = 1000$ to the observation of the time to clearance of worms in a nest. We also investigated the effect of the number of imputations K on the performance of the procedure with $K = 5$, 10 and 500 and obtained similar results. The procedure yielded $\hat{\beta}_1 = 0.0025$ with the estimated standard error of 0.0015 for treatment effect (p-value = 0.0895 for testing of treatment effect) and $\hat{\beta}_2 = -2.2611 \times 10^{-5}$ with the estimated standard error of 5.8895×10^{-5} for age effect (p-value = 0.7011 for testing of age effect). These results suggest that there seems to be no difference in cleaning the worms between the two treatments and the clearance of the worms is not significantly related to the age of the patients at baseline. These results are similar to those given in Li et al. (2012) under the same additive hazards model. However, one must be careful about interpreting the results due to the small number of patients.

6. Discussion and concluding remarks

This paper discussed regression analysis of interval-censored failure time data under the additive hazards model and for the analysis, a multiple imputation approach was presented and investigated. Compared to the existing methods for the problem, a major advantage of the imputation approach is its simplicity. It avoids estimating the correlation of interval-censored observations within the same cluster by using a resampling-based method. The imputation approach converts the regression problem of interval-censored failure time data to that of right-censored failure time data which can be easily implemented by using existing software packages. The numerical studies showed that the imputation approach presented here is efficient and performs well for practical situations.

Compared to the regression approach proposed by Li et al. (2012), the imputation approach is less biased and provides much better coverage probabilities for all the situations considered. The approach by Li et al. (2012) only applies to clustered interval-censored failure time data given in (U_{ij}, V_{ij}) where U and V are two monitoring times assumed to follow a Cox model. However, in real situations, the failure time may only be known to lie within a time interval in the form of $(L_{ij}, R_{ij}]$ with $L \leq R$ and both L and R belonging to $(0, +\infty)$. In this paper, we considered general clustered interval-censored failure time data given by $(L_{ij}, R_{ij}]$ which arise naturally from longitudinal studies with periodic follow-up. The proposed approach can also be conveniently implemented with clustered interval-censored failure time data given by (U_{ij}, V_{ij}). It should be noted that the proposed approach does require estimation of the cumulative baseline hazard function.

There exist several directions for future research. One is that although the simulation suggests that the normal approximation seems reasonable to the distribution of $\hat{\beta}$ and $\hat{\beta}$ is efficient, it would be helpful to provide rigorous justification to it. By looking at the lymphatic filariasis study, one may be also interested in developing a procedure to measure the strength of correlation among subjects within the same cluster and incorporate the correlation in the imputation model. In this paper, we have focused on the situation where covariates are time-independent and subject-specific and it is easy to see that sometimes we

may face time-dependent and/or cluster-specific covariates. Although the proposed approach cannot be directly applied to this latter situation, it can be easily generalized to the situation. Note that the simulation studies have shown that the proposed method works well on the scenarios with moderate to larger number of clusters and small cluster size. Suggested by a reviewer, we also investigated the performance of the method for the scenarios with small number of clusters but large cluster size that may occur in practice, and the results indicate that one needs to be careful for the application of the method to the latter situations and may need to develop a more proper approach. Our simulation results suggest that one needs 50 or more clusters for the presented approach to perform properly.

Another possible direction is that sometimes one may face situations with an informative cluster size in the sense that cluster size is related to the failure time (Williamson et al., 2008). It is apparent that it would be useful to generalize the proposed estimation procedure to the case of clustered interval-censored failure time data with informative cluster sizes under a semiparametric failure time model. In addition, one can see that the additive hazards model (4) assumes a normal distribution for the latent variable b_i and one may want to consider the situation where the covariates may affect the latent variable.

Acknowledgments

The authors would like to thank the co-editor and three reviewers for their very insightful comments and suggestions which greatly improved the paper. Also we want to thank Drs. Gerusa Dreyer and John Williamson for kindly providing the lymphatic filariasis data and Dr. Junlong Li for very helpful discussions. Chengjie Xiong's work was partly supported by the National Institute on Aging (NIA) grant R01 AG034119 as well as by the NIA grant P50 AG05681, P01 AG03991, P01AG26276, and U01 AG032438.

References

Bellamy SL, Li Y, Ryan LM, Lipsitz S, Canner MJ, Wright R. Analysis of clustered and interval censored data from a community-based study in asthma. Stat. Med. 2004; 23:3607–3621. [PubMed: 15534894]

Breslow, NE.; Day, NE. Statistical Methods in Cancer Research, Volume II: The Design and Analysis of Cohort Studies. Oxford University Press for International Agency for Research on Cancer; Oxford: 1987.

Cai J, Prentice RL. Regression estimation using multivariate failure time data and a common baseline hazard function model. Lifetime Data Anal. 1997; 3:197–213. [PubMed: 9384652]

Cox, DR.; Oakes, D. Analysis of Survival Data. Chapman & Hall; London: 1984.

Dreyer G, Addiss D, Williamson J, Norões J. Efficacy of co-administered diethylcarbamazine and albendazole against adult wuchereria bancrofti. Trans. R. Soc. Trop. Med. Hyg. 2006; 100:1118–1125. [PubMed: 16860830]

Hoffman EB, Sen PK, Weinberg CR. Within-cluster resampling. Biometrika. 2001; 88:1121–1134.

Lam KF, Xu Y, Cheung TL. A multiple imputation approach for clustered interval-censored survival data. Stat. Med. 2010; 29:680–693. [PubMed: 20069624]

Li J, Tong X, Sun J. Sieve estimation for the cox model with clustered interval-censored failure time data. Stat. Biosci. 2014; 6:55–72.

Li J, Wang C, Sun J. Regression analysis of clustered interval-censored failure time data with the additive hazards model. J. Nonparametr. Stat. 2012; 24:1041–1050. [PubMed: 25914511]

Lin D, Ying Z. Semiparametric analysis of the additive risk model. Biometrika. 1994; 81:61–71.

Pan W. A multiple imputation approach to cox regression with interval-censored data. Biometrics. 2000; 56:199–203. [PubMed: 10783796]

Rubin, DB. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons; New York: 1987.

Williamson JM, Kim HY, Manatunga A, Addiss DG. Modeling survival data with informative cluster size. Stat. Med. 2008; 27:543–555. [PubMed: 17640035]

Xiang L, Ma X, Yau KK. Mixture cure model with random effects for clustered interval-censored survival data. Stat. Med. 2011; 30:995–1006. [PubMed: 21472759]

Zhang X, Sun J. Regression analysis of clustered interval-censored failure time data with informative cluster size. Comput. Statist. Data Anal. 2010; 54:1817–1823.

Zhang X, Sun J. Semiparametric regression analysis of clustered interval-censored failure time data with informative cluster size. Int. J. Biostat. 2013; 9:205–214. [PubMed: 23940070]

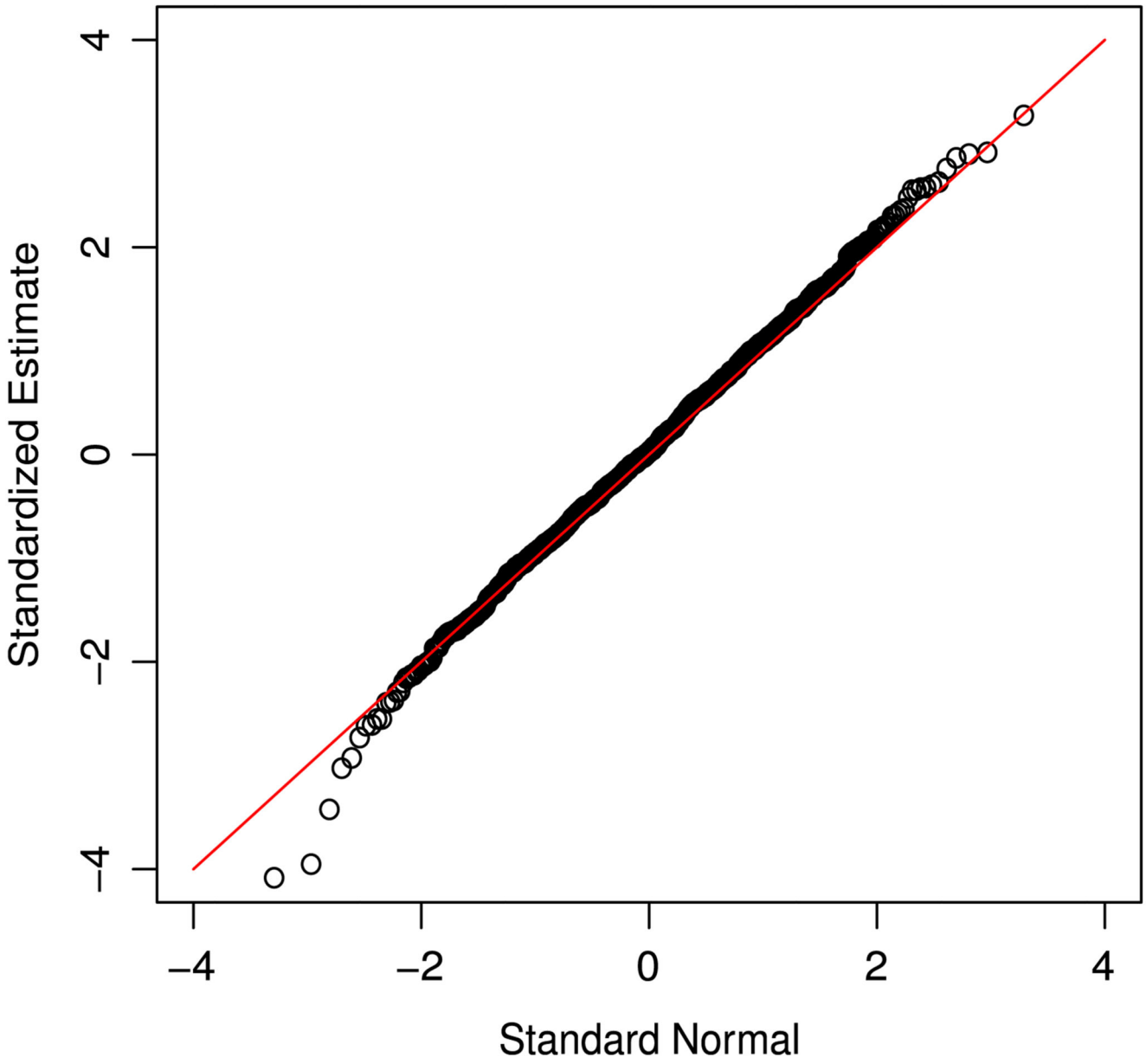

Fig. 1.
Quantile plot of the estimates with covariate-independent censoring intervals ($\gamma_0 = 0$, $\beta_0 = -0.25$, $n = 200$).

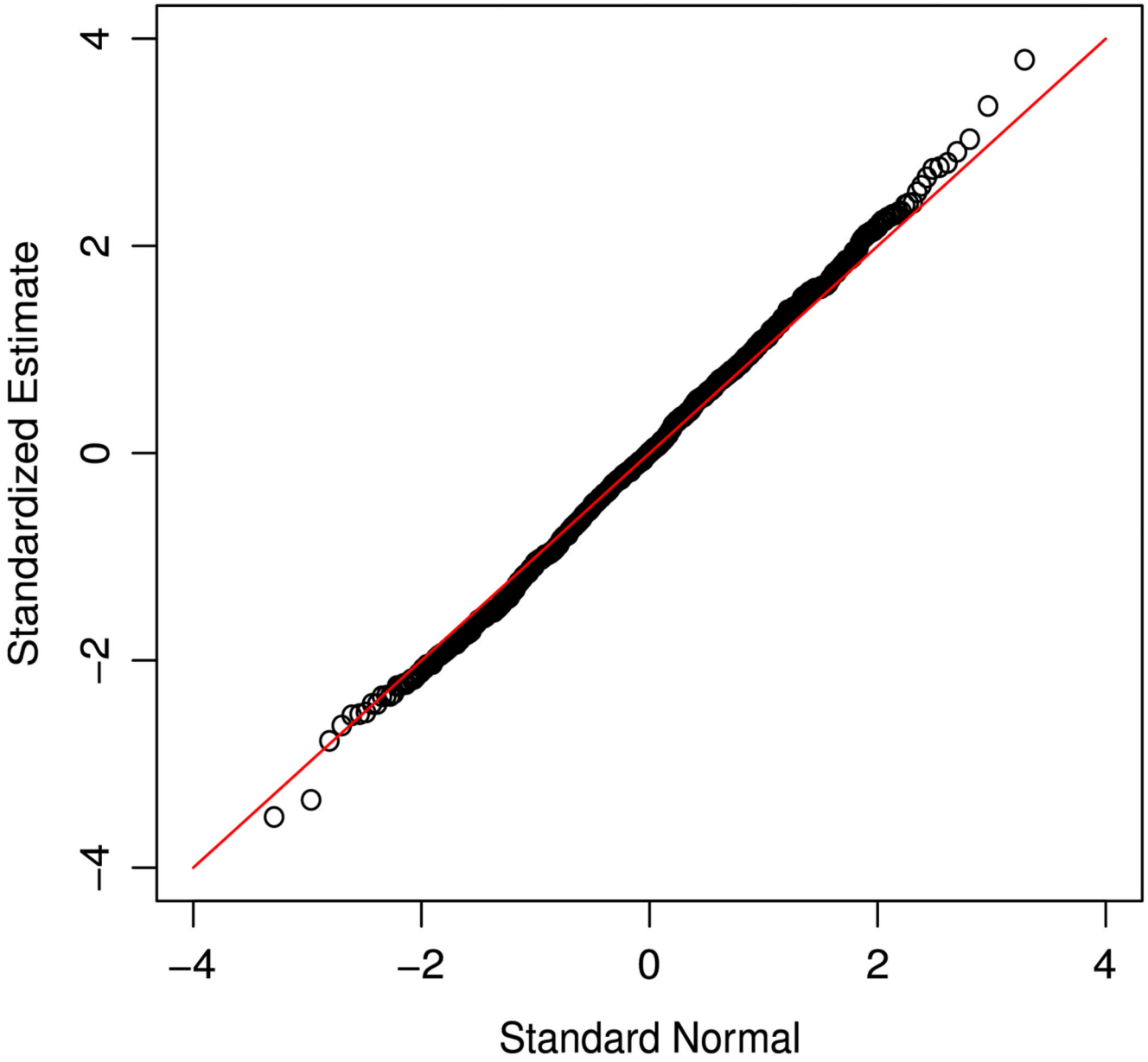

Fig. 2.
Quantile plot of the estimates with covariate-dependent censoring intervals ($\gamma_0 = 0.25$, $\beta_0 = 0.25$, $n = 200$).

Table 1

Estimation of β with $\gamma_0 = 0$ and covariate-independent censoring intervals.

β_0	n	The proposed estimator				Li–Wang–Sun's estimator			
		BIAS	ESD	SSE	CP	BIAS	ESD	SSE	CP
−0.25	100	−0.0134	0.2445	0.2516	0.936	0.0832	0.1779	0.2049	0.894
0	100	−0.0091	0.2585	0.2599	0.942	−0.0008	0.1846	0.2199	0.907
0.25	100	−0.0139	0.2708	0.2911	0.941	−0.1053	0.1916	0.2325	0.860
−0.25	200	−0.0037	0.1744	0.1792	0.943	0.0885	0.1238	0.1422	0.855
0	200	−0.0059	0.1839	0.1773	0.954	0.0053	0.1290	0.1480	0.912
0.25	200	0.0031	0.1932	0.2029	0.943	0.1433	0.1339	0.1564	0.840

Table 2

Estimation of β with $\gamma_0 = -0.25$ and covariate-dependent censoring intervals.

β_0	n	The proposed estimator				Li–Wang–Sun's estimator			
		BIAS	ESD	SSE	CP	BIAS	ESD	SSE	CP
−0.25	100	0.0056	0.2472	0.2521	0.952	0.0967	0.1958	0.2270	0.725
0	100	−0.0011	0.2578	0.2572	0.958	0.1830	0.2077	0.2289	0.835
0.25	100	0.0005	0.2720	0.2755	0.941	0.0964	0.2189	0.2441	0.903
−0.25	200	0.0044	0.1746	0.1715	0.952	0.2886	0.1376	0.1551	0.534
0	200	−0.0002	0.1828	0.1819	0.954	0.1733	0.1461	0.1655	0.766
0.25	200	0.0000	0.1928	0.1979	0.941	0.0947	0.1526	0.1793	0.849

Table 3

Estimation of β with $\gamma_0 = 0.25$ and covariate-dependent censoring intervals.

β_0	n	The proposed estimator				Li–Wang–Sun's estimator			
		BIAS	ESD	SSE	CP	BIAS	ESD	SSE	CP
−0.25	100	−0.0039	0.2565	0.2584	0.945	−0.0968	0.2191	0.2448	0.901
0	100	−0.0135	0.2702	0.2871	0.936	−0.1745	0.2274	0.2579	0.870
0.25	100	−0.0126	0.2859	0.2946	0.939	−0.2820	0.2357	0.2626	0.768
−0.25	200	−0.0016	0.1810	0.1755	0.942	−0.0852	0.1536	0.1642	0.893
0	200	−0.0005	0.1900	0.1883	0.961	−0.1635	0.1592	0.1759	0.815
0.25	200	0.0020	0.2025	0.2081	0.949	−0.2636	0.1655	0.1909	0.641

Table 4

Estimation of β with two covariates and covariate-independent censoring intervals.

β_0	*n*	BIAS	ESD	SSE	CP
$\beta_{10} = 0$	100	−0.0115	0.1971	0.2023	0.942
$\beta_{20} = 0$	100	−0.0062	0.0703	0.0754	0.947
$\beta_{10} = 0.5$	100	0.0093	0.2103	0.2091	0.937
$\beta_{20} = -0.25$	100	0.0039	0.0767	0.0725	0.945
$\beta_{10} = -0.25$	100	−0.0156	0.3142	0.3210	0.941
$\beta_{20} = 0.5$	100	0.0197	0.1406	0.1338	0.930
$\beta_{10} = 0$	200	−0.0066	0.1429	0.1521	0.947
$\beta_{20} = 0$	200	0.0010	0.0675	0.0698	0.951
$\beta_{10} = 0.5$	200	−0.0067	0.1479	0.1470	0.943
$\beta_{20} = -0.25$	200	−0.0013	0.0519	0.0496	0.950
$\beta_{10} = -0.25$	200	−0.0078	0.2393	0.2381	0.946
$\beta_{20} = 0.5$	200	0.0083	0.0941	0.0956	0.938

Table 5

Estimation of β with covariate-dependent censoring intervals.

β_0	n	BIAS	ESD	SSE	CP
−0.25	100	−0.0215	0.3974	0.4202	0.930
0	100	−0.0051	0.4267	0.4724	0.914
0.25	100	0.0067	0.4466	0.4770	0.922
−0.25	200	−0.0106	0.2835	0.2918	0.946
0	200	−0.0021	0.2998	0.3171	0.940
0.25	200	−0.0041	0.3192	0.3272	0.945