Communications for Statistical Applications and Methods
2023, Vol. 30, No. 4, 389–402
https://doi.org/10.29220/CSAM.2023.30.4.389
Print ISSN 2287-7843 / Online ISSN 2383-4757

A GEE approach for the semiparametric accelerated lifetime model with multivariate interval-censored data

Maru Kim[a], Sangbum Choi[1,a]

[a]Department of Statistics, Korea University, Korea

Abstract

Multivariate or clustered failure time data often occur in many medical, epidemiological, and socio-economic studies when survival data are collected from several research centers. If the data are periodically observed as in a longitudinal study, survival times are often subject to various types of interval-censoring, creating multivariate interval-censored data. Then, the event times of interest may be correlated among individuals who come from the same cluster. In this article, we propose a unified linear regression method for analyzing multivariate interval-censored data. We consider a semiparametric multivariate accelerated failure time model as a statistical analysis tool and develop a generalized Buckley-James method to make inferences by imputing interval-censored observations with their conditional mean values. Since the study population consists of several heterogeneous clusters, where the subjects in the same cluster may be related, we propose a generalized estimating equations approach to accommodate potential dependence in clusters. Our simulation results confirm that the proposed estimator is robust to misspecification of working covariance matrix and statistical efficiency can increase when the working covariance structure is close to the truth. The proposed method is applied to the dataset from a diabetic retinopathy study.

Keywords: accelerated failure time model, Buckley-James method, clustered data, generalized estimating equation, interval-censored data, survival analysis

1. Introduction

Survival data, also called time-to-event data often occur in many medical, epidemiological, and socio-economic studies with various types of censoring schemes. Among them, right-censored data have been most frequently studied as a standard survival data format, along with well-established methods and theories in statistics and applied statistics (Kalbfleisch and Prentice, 2002). In practice, however, many clinical studies may involve more complex types of censoring, such as interval-censored data, partly interval-censored data, and doubly interval-censored data, for which existing inference methodologies and theories cannot be used in a straightforward manner. By “interval-censored” data, we usually mean that the failure time of interest cannot be observed directly, but is only known to have occurred within a time interval. Interval-censoring commonly occurs in many areas, such as epidemiological experiments, and medical and longitudinal studies, particularly when subjects cannot be followed up.

In practice, there are several types of interval-censored data. When the failure event is only known to occur within a certain time interval, it is called “case-2” interval-censoring. If there is a single

[1] Corresponding author: Department of Statistics, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea.
E-mail: choisang@korea.ac.kr

Published 31 July 2023 / journal homepage: http://csam.or.kr

examination at a particular time, then the subject is known to have already suffered from the event or not at the time. This is the so-called "case-1" interval-censoring. If failure times are exactly observed for some subjects while others are only known to lie within certain intervals, the dataset is partly interval-censored (PIC) data or doubly-censored (DC) data, depending on the presence of left-censoring. We can easily see these types of data in many clinical and epidemiological researches. For example, the event of interest could be diabetic nephropathy or HIV infection, which is often discovered during periodic clinic visits. If patients can take frequent visits, then their failure times can be ascertained with adequate accuracy. However, since patients can usually take a limited number of visits in practice, their failure times are known to lie within some intervals that may be too broad to be treated as exact. It should be noted that semiparametric estimation methods and associated theories vary considerably depending on the existence of exact failure time observations for analysis of interval censoring data. See Sun (2006) for a comprehensive but thorough examination on this topic.

Regression analysis of interval-censored data with or without exact observations has been extensively studied under various models. Many early studies on interval-censored data focused on the nonparametric and semiparametric maximum likelihood estimation of underlying distribution function (Huang, 1996; Huang and Wellner, 1997; Wellner and Zhan, 1997; Shen, 1998). Estimation of survival function via self-consistency equation was studied by Turnbull (1976). Huang (1999) studied nonparametric estimation of distribution function based on PIC data, and Kim (2003) studied the nonparametric maximum likelihood estimation (NPMLE) for proportional hazards regression models. Zhao *et al*. (2008) presented a class of generalized log-rank tests for PIC data and established their asymptotic properties. Several *M*-estimation methods were investigated for linear regression with PIC or DC data, assuming that the censoring time is assumed to be completely independent of event time (Ren and Gu, 1997; Ren, 2003). Lin *et al*. (2012) and Ji *et al*. (2012) alternatively examined censored quantile regression models for DC data, based on the idea of redistribution-of-mass (Portnoy, 2003) and martingale equation Peng and Huang (2008), respectively. Kim *et al*. (2010) extended the missing information principle (Efron, 1967) to the interval-censored median regression, while the covariate type is restrictive as the continuous type is not affordable. More recently, a broad class of semiparametric transformation models (Zeng *et al*., 2016; Mao *et al*., 2017; Gao and Chan, 2019; Choi and Huang, 2021) and accelerated failure time (AFT) model (Gao *et al*., 2017; Choi *et al*., 2021) were considered to analyze various types of interval-censored data.

Thus far, it is assumed that the event of interest occurs independently for each individual. However, failure time data are often clustered, where the subjects from the same cluster may be correlated, but the subjects in different clusters can be treated to be independent. Several statistical methods have been suggested to make inferences with clustered interval-censored data. Lam *et al*. (2010) proposed a simple multiple imputation strategy to recover the order of occurrences based on the interval-censored event times. Kor *et al*. (2013) discussed regression analysis of clustered interval-censored data based on Cox's proportional hazards model. Chen *et al*. (2016) considered the same problem but under the semiparametric additive hazards mode by using a multiple imputation approach for inference. Zeng *et al*. (2017) investigated the effects of possibly time-dependent covariates on multivariate failure times by considering a linear transformation model with random effects and developed a novel nonparametric maximum likelihood estimation under general interval-censoring schemes. However, most approaches left the within-cluster dependence structure unspecified or just assumed independence. Moreover, little work has been done to extend the generalized estimating equation (GEE) approach to accommodate the dependence of clustered interval-censored data on AFT models.

In this article, we propose the Buckley and James (1979) method for imputing clustered interval-censored data and develop an iterative GEE procedure for the AFT model that relates the logarithm of the failure time linearly to the covariates. We explore a general Buckley-James imputation method for multivariate case-1, case-2 and partly interval-censored data. The proposed Buckley-James (BJ) method enables us to impute the latent event time by its estimable conditional expectation given covariate. Since the imputed pseudo response involves the unknown residual distribution function, we estimate it by using a modified self-consistency equation (Choi *et al*., 2021) After approximating the pseudo-response of the latent event time through an iterative BJ procedure, we use a GEE procedure to estimate the regression coefficient while accounting for potential correlation structure in clusters. This method has the same concept as a GEE approach for complete data, because misspecification of the working covariance matrix does not affect the consistency of the parameter estimator in the multivariate linear model. When the working covariance matrix is close to the unknown true working covariance matrix, the estimator has more statistical efficiency than that from the working independence assumption (Chiou *et al*., 2014).

The remainder of this paper is organized as follows. The semiparametric AFT model and data are introduced in Section 2.1 and methods to deal with clustered interval-censored data are explained in Section 2.2. In Section 2.3, we propose the GEE approach with details. Section 3 reports simulation results to assess the finite-sample properties of the proposed estimator. In Section 4, our proposed methods are applied to the real data. We close the article with a discussion in Section 5.

2. Methods

2.1. Data and model

Consider multivariate failure time data that consist of n clusters and K_i subjects within cluster $i = 1, \ldots, n$. For simplicity, we assume all n clusters have the same cluster size K, i.e., $K_i = K$. For $i = 1, \ldots, n$ and $k = 1, \ldots, K$, let T_{ik} be the failure time for the k^{th} subject in the i^{th} cluster, and $X_i = (X_{i1}, \ldots, X_{iK})$ be the $p \times K$ covariate matrix of the i^{th} cluster, with the k^{th} column being denoted by X_{ik}. The semiparametric multivariate AFT model for the k^{th} subject in the i^{th} cluster specifies

$$\log T_{ik} = \beta' X_{ik} + \varepsilon_{ik}, \quad (i = 1, \ldots, n,\ k = 1, \ldots, K), \tag{2.1}$$

where β is a p-vector of regression coefficients and $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iK})$ is a random vector, independent and identically distributed across clusters with unknown underlying distribution function F. Within a cluster, the components of $\varepsilon_{i1}, \ldots, \varepsilon_{iK}$ do not need to follow a common distribution and may be correlated.

Under a general clustered interval-censoring scheme, we can formulate clustered interval-censored data by considering a random sequence of examination times, denoted by $0 = U_{ik0} < U_{ik1} < \cdots < U_{ikQ} < U_{ik(Q+1)} = \infty$, where Q is a number of examination times for the k^{th} subject in the i^{th} cluster. We do not need to model the entire sequence of these examination times, but can focus on the smallest observed interval (L_{ik}, R_{ik}) that contains the latent event time T_{ik}, because (L_{ik}, R_{ik}) only contributes to the likelihood construction. Here, $L_{ik} = \max\{U_{ikq} : U_{ikq} \le T_{ik},\ q = 0, \ldots, Q\}$ and $R_{ik} = \min\{U_{ikq} : U_{ikq} > T_{ik},\ q = 1, \ldots, Q+1\}$ for the k^{th} subject in the i^{th} cluster. That is, (L_{ik}, R_{ik}) represents the tightest observed interval that contains T_{ik}.

Since there are several different types of clustered interval-censored data in practice and their corresponding censoring mechanism is formalized in different ways, we next list the most relevant cases. When $Q \ge 2$, namely, in experiments with multiple examination times per subject, we only know the survival time of interest T_{ik} has occurred (i) before the left end point of the time interval ($T_{ik} \le L_{ik}$), or (ii) within some random time interval ($L_{ik} < T_{ik} \le R_{ik}$), or (iii) after the right end point of the time interval ($T_{ik} > R_{ik}$). When $Q = 1$, that is, there is a one-time examination per each subject (let U_{ik} be the single examination time), case-1 interval-censored (or current status) data are created, in which the only knowledge about the failure time T_{ik} is whether it has occurred before U_{ik} or not. In consequence, the survival time is either left-censored or right-censored. If a non-ignorable proportion of exact failure times is also available for some patients in addition to these interval-censoring sampling schemes, PIC or DC data are observed. That is, case-2 interval-censored data plus exact observations are equivalent to PIC data, whereas case-1 interval-censored data plus exact observations are DC data.

2.2. Generalized Buckley-James imputation for interval-censored data

2.2.1. Clustered case-2 interval-censored data

We first describe a new BJ method to impute censored failure times under clustered case-2 interval-censoring. Let $Y_{ik} = \log T_{ik}$ be the log-transformed failure time for the k^{th} subject in the i^{th} cluster. Since Y_{ik} is interval-censored, we estimate its conditional expectation, given that $L_{ik} < T_{ik} < R_{ik}$ and X_{ik}, under model (2.1). By modifying the BJ imputation method for right-censored data, we define the pseudo response of Y_{ik}, denoted by Y^*_{ik}, as

$$Y^*_{ik}(\beta) = E(Y_{ik} \mid L_{ik} < T_{ik} < R_{ik}, X_{ik}) = \beta' X_{ik} + E(\varepsilon_{ik} \mid l_{ik}(\beta) < \varepsilon_{ik} < r_{ik}(\beta))$$
$$= \beta' X_{ik} + \frac{\int_{l_{ik}(\beta)}^{r_{ik}(\beta)} u dF(u)}{F\{r_{ik}(\beta)\} - F\{l_{ik}(\beta)\}}, \tag{2.2}$$

where $l_{ik}(\beta) = \log L_{ik} - \beta' X_{ik}$ and $r_{ik}(\beta) = \log R_{ik} - \beta' X_{ik}$ denote the residual terms, corresponding to the endpoints in (L_{ik}, R_{ik}), respectively. Here, we assume that K clusters share a common residual distribution F, because the cluster-specific distribution function is difficult to obtain from the current setting. Notice that equation (2.2) assumes that F is known, which in fact should be estimated from the data.

In order to estimate the residual distribution function F, one may solve the self-consistency equation (Turnbull, 1974, 1976) iteratively until convergence:

$$F(t) = \frac{1}{nK} \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{F\{r_{ik}(\beta) \wedge t\} - F\{l_{ik}(\beta) \wedge t\}}{F\{r_{ik}(\beta)\} - F\{l_{ik}(\beta)\}}. \tag{2.3}$$

See also Wellner and Zhan (1997) for a more efficient iterative convex minorant algorithm to approximate F. However, these equation-based methods are usually difficult to implement and do not necessarily result in the nonparametric maximum likelihood estimator (NPMLE) for F. Therefore, we alternatively used a novel modified expectation-maximization (EM) algorithm (Choi *et al.*, 2021) to maximize the nonparametric log-likelihood function for F given β, based on $\{(l_{ik}(\beta), r_{ik}(\beta)), i = 1, \ldots, n, k = 1, \ldots, K\}$.

To be specific, let $\Lambda(\cdot) \equiv \Lambda(\cdot, \beta) = -\log\{1 - F(\cdot, \beta)\}$ denote the baseline cumulative hazards function of the error term. Let $-\infty = s_0 < s_1 < \cdots < s_m < \infty$ denote the unique and ordered event times of the observed set of $\{(L_{ik}, R_{ik}); i = 1, \ldots, n, k = 1, \ldots K\}$, where m is the number of unique time points in a finite time horizon. With a slight abuse of notation, we write $d\Lambda_j \equiv d\Lambda(s_j)$ to denote the jump size of $\Lambda(\cdot)$ at time s_j and let $\Lambda_j = \Lambda(s_j) \equiv \sum_{t \le s_j} d\Lambda(t)$. To invoke the EM algorithm, let $\{W_{ikj}\}$ be independent subject-specific Poisson random variables with mean $\mu = d\Lambda_j$ with density function $p(W = w; \mu) = e^{-\mu}\mu^w / w!$. The nonparametric log-likelihood function for $\Lambda(\cdot)$ for a fixed β

can be written as

$$\ell(\Lambda) = \sum_{i=1}^{n}\sum_{k=1}^{K} \log\left[\exp\{-\Lambda(l_{ik})\} - I(R_{ik} < \infty)\exp\{-\Lambda(r_{ik})\}\right]$$
$$= \sum_{i=1}^{n}\sum_{k=1}^{K}\left[\log P\left(\sum_{j:s_j\le L_{ik}} W_{ikj} = 0\right) + I(R_{ik} < \infty)\log\left\{1 - P\left(\sum_{j:L_{ik}<s_j\le R_{ik}} W_{ikj} > 0\right)\right\}\right].$$

Assuming that the W_{ikj}'s are known, the complete-data log-likelihood function takes a simple form $\ell^c(\Lambda; W) = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{m} p(W_{ikj}, d\Lambda_j)$ under the constraints that $\sum_{j:s_j\le L_{ik}} W_{ikj} = 0$ and $I(R_{ik} < \infty)\sum_{j:L_{ik}<s_j\le R_{ik}} W_{ikj} > 0$. The E-step shows that the expected complete-data log-likelihood function is equivalent to

$$E_W\{\ell^c(\Lambda; W)\} = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{m} I\left(s_j \le \tilde{R}_{ik}\right)\left\{\log\left(d\Lambda_j\right)\xi_{ikj} - d\Lambda_j\right\},$$

where $\tilde{R}_{ik} = R_{ik}I(R_{ik} < \infty) + L_{ik}I(R_{ik} = \infty)$ and the conditional expectation $\xi_{ikj} \equiv E_W(W_{ikj})$ is given by

$$\xi_{ikj} = \left\{\frac{d\Lambda_j I\left(L_{ik} < s_j \le R_{ik} < \infty\right)}{1 - \exp\left(-\sum_{j:L_{ik}<s_j\le R_{ik}} d\Lambda_j\right)}\right\} + d\Lambda_j I\left(s_j > \tilde{R}_{ik}\right). \tag{2.4}$$

In the M-step, we can obtain a closed-form expression for $d\Lambda_j$ by solving the score likelihood equation with respect to $d\Lambda_j$, which leads to

$$d\hat{\Lambda}_j = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K}\xi_{ikj} I\left(s_j \le \tilde{R}_{ik}\right)}{\sum_{i=1}^{n}\sum_{k=1}^{K} I\left(s_j \le \tilde{R}_{ik}\right)}, \quad j = 1, \ldots, m. \tag{2.5}$$

Therefore, the proposed EM-based NPMLE for $\Lambda(\cdot)$ can then be calculated by simply iterating the E-step (2.4) and the M-step (2.5) until convergence. The resulting estimator for $F(\cdot)$ is then obtained by $\hat{F}(\cdot) = 1 - \exp\{-\hat{\Lambda}(\cdot)\}$. Once we obtain the NPMLE $\hat{F}$ for F, we can use $\hat{Y}_{ik}(\beta)$ to approximate $Y^*_{ik}(\beta)$, where

$$\hat{Y}_{ik}(\beta) = \beta' X_{ik} + \frac{\int_{l_{ik}(\beta)}^{r_{ik}(\beta)} u d\hat{F}(u)}{\hat{F}\{r_{ik}(\beta)\} - \hat{F}\{l_{ik}(\beta)\}}. \tag{2.6}$$

Equation (2.6) implies that the pseudo response $\hat{Y}_{ik}(\beta)$ is a weighted average of failure times over $[l_{ik}(\beta), r_{ik}(\beta)]$, based on the estimated residual function $\hat{F}$. We note that when interval-censoring is replaced by right-censoring, equation (2.6) reduces to the standard BJ imputation for right-censored data. Therefore, our imputation strategy generalizes the conventional BJ method to more complex censoring schemes.

2.2.2. Clustered case-1 interval-censored data

We next consider the proposed BJ imputation method for clustered case-1 interval-censoring (current status) data, in which we observe $\{(C_{ik}, \delta_{ik}, X_{ik}), i = 1, \ldots, n, k = 1, \ldots, K\}$, where C_{ik} denotes the single examination time and $\delta_{ik} = I(T_{ik} \le C_{ik})$ is the censoring indicator. Note that current status data is a mixture of left-censoring ($\delta_{ik} = 1$) and right-censoring ($\delta_{ik} = 0$). Let the observed error term under model (2.1) denote by $e_{ik}(\beta) = \log C_{ik} - \beta' X_{ik}$. Since case-1 interval-censoring is a special case of case-2 interval-censoring with $L_{ik} = -\infty$ or $R_{ik} = \infty$, the imputed pseudo-response can be approximated by

$$\hat{Y}_{ik}(\beta) = \beta' X_{ik} + \delta_{ik} \frac{\int_{-\infty}^{e_{ik}(\beta)} u d\hat{F}(u)}{\hat{F}\{e_{ik}(\beta)\}} + (1 - \delta_{ik}) \frac{\int_{e_{ik}(\beta)}^{\infty} u d\hat{F}(u)}{1 - \hat{F}\{e_{ik}(\beta)\}}. \tag{2.7}$$

To estimate the residual function F, one may solve the following self-consistency equation

$$F(t) = \frac{1}{nK} \sum_{i=1}^{n} \sum_{k=1}^{K} \left[\delta_{ik} \frac{F\{e_{ik}(\beta) \wedge t\}}{F\{e_{ik}(\beta)\}} + (1 - \delta_{ik}) \left\{ \frac{F(t) - F\{e_{ik}(\beta) \wedge t\}}{1 - F\{e_{ik}(\beta)\}} \right\} \right], \tag{2.8}$$

or use a similar EM algorithm described above. Note that two formulations (2.7) and (2.8) are implied from (2.6) and (2.3) respectively by letting $F(l_{ik}) = F(-\infty) = 0$ if $\delta_{ik} = 1$ and $F(r_{ik}) = F(\infty) = 1$ if $\delta_{ik} = 0$.

2.2.3. Clustered partly interval-censored data

Our main focus in this article is clustered partly interval-censored (PIC) data. In this situation, the failure times are exactly observed for some subjects, but only known to be within a certain time interval for the rest. For clustered PIC data, we can observe $\{(\delta_{ik}, \delta_{ik} T_{ik}, (1-\delta_{ik}) L_{ik}, (1-\delta_{ik}) R_{ik}, X_{ik}), i = 1, \ldots, n, k = 1, \ldots, K\}$, where $\delta_{ik} = I(L_{ik} = R_{ik})$ is the censoring indicator. PIC data is in fact a mixture of exact observations (i.e., $\delta_{ik} = 1$) and interval-censored data (i.e., $\delta_{ik} = 0$). Therefore, we can derive the pseudo response of Y_{ik}, denoted by Y_{ik}^*, as

$$\begin{aligned} Y_{ik}^*(\beta) &= \delta_{ik} Y_{ik} + (1 - \delta_{ik}) E(Y_{ik} \mid L_{ik} < T_{ik} < R_{ik}, X_{ik}) \\ &= \delta_{ik} Y_{ik} + (1 - \delta_{ik}) [\beta' X_{ik} + E\{\varepsilon_{ik} \mid l_{ik}(\beta) < \varepsilon_{ik} < r_{ik}(\beta)\}] \\ &= \delta_{ik} Y_{ik} + (1 - \delta_{ik}) \left[\beta' X_{ik} + \frac{\int_{l_{ik}(\beta)}^{r_{ik}(\beta)} u dF(u)}{F\{r_{ik}(\beta)\} - F\{l_{ik}(\beta)\}} \right]. \end{aligned} \tag{2.9}$$

We note that under conventional right-censored data, the above imputation formula reduces to the standard BJ equation:

$$Y_{ik}^*(\beta) = \delta_{ik} Y_{ik} + (1 - \delta_{ik}) \left[\beta' X_{ik} + \frac{\int_{e_{ik}(\beta)}^{\infty} u dF(u)}{1 - F\{e_{ik}(\beta)\}} \right],$$

where $e_{ik}(\beta) = Y_{ik} - \beta' X_{ik}$ represents the observed residual term under model (2.1). Equation (2.9) can also include the BJ method for DC data by letting $l_{ik}(\beta) = -\infty$ for left-censoring and $r_{ik}(\beta) = \infty$ for right-censoring.

To approximate Y_{ik}^*, we again have to estimate F in a similar way, as discussed in Section 2.2.1. To this end, one may solve iteratively the self-consistency equation

$$F(t) = \frac{1}{nK} \sum_{i=1}^{n} \sum_{k=1}^{K} \left[\delta_{ik} I\{e_{ik}(\beta) \le t\} + (1 - \delta_{ik}) \left\{ \frac{F\{r_{ik}(\beta) \wedge t\} - F\{l_{ik}(\beta) \wedge t\}}{F\{r_{ik}(\beta)\} - F\{l_{ik}(\beta)\}} \right\} \right] \tag{2.10}$$

until convergence or develop an EM-based method to locate an NPMLE. After we get $\hat{F}$, we can make use of $\hat{Y}_{ik}(\beta)$ to approximate Y^*_{ik}.

$$\hat{Y}_{ik}(\beta) = \delta_{ik}Y_{ik} + (1-\delta_{ik})\left[\beta' X_{ik} + \frac{\int_{l_{ik}(\beta)}^{r_{ik}(\beta)} u d\hat{F}(u)}{\hat{F}\{r_{ik}(\beta)\} - \hat{F}\{l_{ik}(\beta)\}}\right]. \tag{2.11}$$

2.3. GEE-based inference procedure

Once we obtain the pseudo response for each individual, we can define $\hat{Y}_i(\beta) = \{\hat{Y}_{i1}(\beta), \ldots, \hat{Y}_{iK}(\beta)\}$ for the response vector of the i^{th} cluster. We already defined X_i as a $p \times K$ covariate matrix for cluster i. The BJ estimator (Buckley and James, 1979; Jin *et al.*, 2006) is based on the following least-square operator

$$L_n(b) = \left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)'\right]^{-1}\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(\hat{Y}_i(b) - \bar{Y}(b)\right)'\right], \tag{2.12}$$

where $\bar{Y}(b) = n^{-1}\sum_{i=1}^{n}\hat{Y}_i(b)$, $\bar{X} = n^{-1}\sum_{i=1}^{n}X_i$ and b is an estimator of β. Equation (2.12) leads to an iterative algorithm:

$$\hat{\beta}_n^{(m)} = L_n\left(\hat{\beta}_n^{(m-1)}\right), \quad (m \geq 1).$$

It is well known that if the initial estimator $\hat{\beta}_n^{(0)}$ is consistent and asymptotically normal, $\hat{\beta}_n^{(m)}$ is consistent and asymptotically normal for every $m \geq 1$ (Jin *et al.*, 2006; Chiou *et al.*, 2014). Although this estimator is consistent, we can improve its efficiency since it completely ignores the within-cluster dependence. Equation (2.12) is a special case of our GEE estimator with a proper working independence covariance structure. Therefore, we will incorporate the within-cluster dependence by using a GEE-based method as follows.

The GEE approach was initially developed by Liang and Zeger (1986) in order to produce regression estimates when analyzing repeated measures with non-normal response variables. For non-censored data, the GEE approach can increase the statistical efficiency of the marginal regression coefficient estimator by incorporating an inverse working covariance matrix as weight into the estimating equation. A working covariance matrix does not need to be correctly specified. However, when a working covariance matrix is close to the truth, it can improve the estimator's efficiency. It may contain additional working parameters, whose estimation does not affect the estimator's consistency. There are several assumptions for this. Observations within a cluster may be correlated but observations in separate clusters should be independent. These assumptions are the same as that in our model. That is why we can use the GEE approach to our AFT modeling which explains dependence by using a working correlation structure to improve our estimator's efficiency. This approach was first adopted by Chiou *et al.* (2014) for longitudinal right-censored survival data. Our purpose here is to extend their approach to various types of interval-censored data.

Suppose now that the working covariance matrix Ω has a parameter vector α. For a given prior estimator b of β, and let $\alpha(b)$ be an estimator of α, we suggest updating the estimator iteratively by solving the following GEE equation:

$$U_n(\beta, b, \alpha) = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)\Omega_i^{-1}\{\alpha(b)\}\left(\hat{Y}_i(b) - \beta' X_i\right)' = 0, \tag{2.13}$$

Table 1: Summary of simulation results with identical regression coefficients and identical marginal error normal distribution for clustered PIC data based on 1000 replications

Error	τ	Cens		Par	Bias			SSE			ASE		
		Exact	π_I		IND	EX	AR1	IND	EX	AR1	IND	EX	AR1
Normal	0.3	80%	20%	β_1	−0.003	−0.006	−0.002	0.069	0.067	0.066	0.071	0.070	0.069
				β_2	0.002	−0.005	−0.006	0.110	0.107	0.103	0.107	0.097	0.101
		50%	50%	β_1	−0.013	−0.006	−0.008	0.070	0.073	0.069	0.072	0.080	0.060
				β_2	0.017	0.085	0.013	0.115	0.111	0.113	0.113	0.116	0.102
		20%	80%	β_1	0.016	0.019	0.014	0.084	0.082	0.079	0.095	0.089	0.084
				β_2	−0.019	−0.018	−0.018	0.126	0.123	0.120	0.130	0.129	0.113
	0.6	80%	20%	β_1	0.001	0.002	0.003	0.069	0.066	0.062	0.073	0.076	0.069
				β_2	0.002	0.001	0.002	0.120	0.117	0.111	0.117	0.113	0.118
		50%	50%	β_1	−0.009	−0.003	−0.008	0.075	0.070	0.069	0.079	0.080	0.075
				β_2	0.003	0.013	0.004	0.126	0.120	0.119	0.127	0.115	0.108
		20%	80%	β_1	0.019	0.021	0.020	0.091	0.086	0.080	0.088	0.093	0.090
				β_2	−0.018	−0.031	−0.022	0.138	0.132	0.127	0.142	0.138	0.130

*π_I is interval-censoring rate; SSE is the sampling standard errors; ASE is the average of estimated standard errors.

where $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$, and Ω_i is a $K \times K$ nonsingular working weight matrix. This working weight matrix Ω_i involves a working parameter α, which may also depend on b. For given α and b, the solution to (2.13) has a closed-form expression

$$L_n(b,\alpha) = \left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\Omega_i^{-1}\{\alpha(b)\}\left(X_i - \bar{X}\right)'\right]^{-1}\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\Omega_i^{-1}\{\alpha(b)\}\left(\hat{Y}_i(b) - \bar{Y}(b)\right)'\right], \quad (2.14)$$

where $\bar{Y}(b) = n^{-1}\sum_{i=1}^{n}\hat{Y}_i(b)$.

The proposed estimation procedure can be carried out iteratively, as summarized as follows:

Step 1 : Obtain an initial estimate $\hat{\beta}_n^{(0)} = b_n$ of β with the naive GEE model and set $m = 0$.

Step 2 : Given $\hat{\beta}_n^{(m-1)}$, update the pseudo response $\hat{Y}_{ik}(\beta)$ using the proposed BJ algorithm.
Obtain an estimate of α, given $\hat{\beta}_n^{(m-1)}$, $\hat{\alpha}_n(\hat{\beta}_n^{(m-1)})$.

Step 3 : Update the regression coefficient $\hat{\beta}_n^{(m)}$ by

$$\hat{\beta}_n^{(m)} = L_n\left(\hat{\beta}_n^{(m-1)}, \hat{\alpha}_n\left(\hat{\beta}_n^{(m-1)}\right)\right).$$

Step 4 : Increase m by 1 and repeat Steps 2 and 3 until convergence.

We stop the iteration when either $\|\hat{\beta}_{(k)} - \hat{\beta}_{(k-1)}\| \leq 10^{-4}$ or the maximum number of iterations (set to 100) is first achieved. When we compare the closed-form of the solution of the independent least-squares equation (2.14) to that of equation (2.12), they are very similar except that equation (2.14) has a working weight matrix $\Omega_i^{-1}(\alpha(b))$. This part enables us to accommodate the within-cluster dependence.

We can achieve the highest efficiency when $\Omega_i(\alpha)$ is chosen to be the true covariance of $Y_i(\beta_0)$. The variance of the estimator can be estimated by a resampling procedure. As we assume that all clusters have the same size K, the working covariance matrix Ω_i's have the same size; they only differ with the i^{th} cluster when the cluster sizes are not equal. In practice, we may impose exchangeable (EX), autoregressive with order one (AR1), or independence (IND) structure as a parsimonious working

Table 2: Summary of simulation results with identical regression coefficients and identical marginal error EV distribution for clustered PIC data based on 1000 replications

Error	τ	Cens		Par	Bias			SSE			ASE		
		Exact	π_I		IND	EX	AR1	IND	EX	AR1	IND	EX	AR1
EV	0.3	80%	20%	β_1	−0.011	−0.013	−0.011	0.046	0.044	0.044	0.044	0.054	0.049
				β_2	0.015	0.009	0.012	0.071	0.071	0.070	0.068	0.061	0.065
		50%	50%	β_1	0.018	−0.019	−0.018	0.048	0.046	0.046	0.044	0.052	0.050
				β_2	0.018	0.020	0.021	0.076	0.074	0.072	0.072	0.065	0.063
		20%	80%	β_1	0.006	0.003	0.004	0.052	0.051	0.050	0.056	0.055	0.046
				β_2	−0.005	−0.003	−0.003	0.078	0.076	0.074	0.081	0.085	0.077
	0.6	80%	20%	β_1	−0.008	−0.012	−0.011	0.049	0.043	0.042	0.047	0.049	0.045
				β_2	0.013	0.016	0.013	0.074	0.067	0.067	0.070	0.060	0.061
		50%	50%	β_1	−0.021	−0.026	−0.020	0.053	0.047	0.048	0.057	0.050	0.051
				β_2	0.024	0.025	0.022	0.081	0.076	0.074	0.079	0.082	0.071
		20%	80%	β_1	0.003	0.006	0.007	0.060	0.054	0.052	0.065	0.063	0.061
				β_2	0.001	−0.007	−0.005	0.087	0.083	0.082	0.091	0.086	0.087

*π_I is interval-censoring rate; SSE is the sampling standard errors; ASE is the average of estimated standard errors.

Table 3: Summary of simulation results with identical regression coefficients and identical marginal error gamma distribution for clustered PIC data based on 1000 replications

Error	τ	Cens		Par	Bias			SSE			ASE		
		Exact	π_I		IND	EX	AR1	IND	EX	AR1	IND	EX	AR1
Gamma	0.3	80%	20%	β_1	−0.012	−0.010	−0.011	0.051	0.050	0.048	0.046	0.058	0.045
				β_2	0.014	0.013	0.013	0.073	0.073	0.071	0.068	0.080	0.063
		50%	50%	β_1	−0.021	−0.020	−0.020	0.052	0.053	0.050	0.054	0.057	0.053
				β_2	0.024	0.025	0.023	0.075	0.075	0.073	0.078	0.079	0.075
		20%	80%	β_1	−0.000	0.001	0.002	0.059	0.056	0.054	0.055	0.057	0.057
				β_2	0.000	0.002	0.001	0.084	0.080	0.080	0.088	0.086	0.086
	0.6	80%	20%	β_1	−0.010	−0.009	−0.012	0.057	0.051	0.047	0.054	0.056	0.052
				β_2	0.016	0.013	0.011	0.075	0.074	0.068	0.079	0.077	0.071
		50%	50%	β_1	−0.021	−0.021	−0.052	0.056	0.051	0.048	0.060	0.062	0.053
				β_2	0.025	0.023	0.024	0.077	0.079	0.072	0.074	0.071	0.069
		20%	80%	β_1	0.001	0.002	0.000	0.058	0.054	0.052	0.062	0.051	0.055
				β_2	0.003	−0.001	−0.001	0.087	0.082	0.081	0.089	0.085	0.078

*π_I is interval-censoring rate; SSE is the sampling standard errors; ASE is the average of estimated standard errors.

covariance. It should be noted that GEE is intended for simple clustering or repeated measures. It cannot easily accommodate more complex designs such as nested or crossed groups; for example, nested repeated measures within a subject or group. This is something better suited for a mixed-effect model (McCulloch and Searle, 2004).

3. Simulation studies

We conducted two simulation studies to assess the finite-sample performance of the proposed BJ estimators with three working covariance structures, i.e., exchangeable (EX), autoregressive with order one (AR1), and independence (IND) structures. In our first simulation study, we assumed a clustered PIC failure time setting with identical regression coefficients across different clusters and identical marginal error distribution. We considered $n = 100$ clusters and each cluster size was fixed at three, i.e., $K = 3$. For cluster i, the failure time $T_i = (T_{i1}, T_{i2}, T_{i3})'$ was generated from

$$\log T_{ik} = X_{1ik} - X_{2ik} + \varepsilon_{ik},$$

Table 4: Summary of simulation results with identical regression coefficients and identical marginal error normal distribution for clustered PIC data based on 1000 replications

Error	τ	Cens		Par	Bias			SSE			ASE		
		Exact	π_I		IND	EX	AR1	IND	EX	AR1	IND	EX	AR1
Normal	0.3	80%	20%	β_1	−0.007	0.001	−0.005	0.046	0.044	0.041	0.062	0.059	0.057
				β_2	0.005	−0.002	−0.003	0.098	0.097	0.095	0.097	0.092	0.091
		50%	50%	β_1	−0.003	−0.003	−0.002	0.065	0.063	0.060	0.062	0.061	0.058
				β_2	0.007	0.063	0.007	0.103	0.099	0.095	0.103	0.102	0.102
		20%	80%	β_1	0.006	0.014	0.002	0.074	0.071	0.069	0.085	0.079	0.078
				β_2	−0.009	−0.008	−0.011	0.116	0.112	0.110	0.119	0.119	0.111
	0.6	80%	20%	β_1	0.004	0.003	0.001	0.063	0.056	0.051	0.063	0.056	0.054
				β_2	0.005	0.002	0.003	0.110	0.090	0.072	0.106	0.092	0.088
		50%	50%	β_1	−0.002	−0.001	−0.003	0.055	0.050	0.039	0.068	0.060	0.055
				β_2	0.005	0.008	0.002	0.115	0.098	0.091	0.137	0.105	0.093
		20%	80%	β_1	0.009	0.011	0.012	0.081	0.075	0.060	0.082	0.080	0.054
				β_2	−0.020	−0.011	−0.012	0.127	0.122	0.107	0.121	0.108	0.101

*π_I is interval-censoring rate; SSE is the sampling standard errors; ASE is the average of estimated standard errors.

where $X_{1ik} \sim N(0,1)$, X_{2ik} can take three distinct numbers $(1,2,3)$ with probability $1/3$ for each. The joint distribution of $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})'$ was specified by a common marginal distribution. Three marginal error distributions were considered: (i) standard normal distribution, $N(0,1)$, (ii) extreme value (EV) distribution, EV(location $= -0.5$, scale $= 0.5$), and (iii) log-transformed gamma distribution, $\log\Gamma$(shape $= 1.5$, scale $= 1$). Dependence within cluster was created via Cholesky factorization, so that Kendall's τ was set to 0.3 and 0.6. We generated two examination times $L \sim$ Uniform$(0,c)$ and $R \sim L +$ Uniform$(0, c+1)$, where the constant c was varied to achieve desired censoring rates. For PIC data, subjects could be left-censored at L, or right-censored at R, or interval-censored in (L,R). By this set-up, the constant c was tuned to achieve the three levels of exact data percentage: 80%, 50%, and 20%. Then three working covariance cases were used: IND, EX and AR1. To evaluate the performance of our method, we used the sample standard error (SSE) and average standard error (ASE). They were based on 1000 replicates with 200 bootstrapped samples for each configuration. To account for the clustered structure, the bootstrap sampling was conducted at the cluster level. That is, once a cluster is sampled, then the entire elements in the cluster will be used for inference.

The simulation results are summarized in Tables 1–3, respectively, for different marginal distributions. All estimators appear to be virtually unbiased. We can confirm the fact that our regression coefficient estimator is consistent and robust to the misspecification of the working covariance. In general, the sample standard error (SSE) and average standard error (ASE) are well-matched, suggesting that the resampling procedure provides a valid inference. For a given censoring percentage, as the dependence level increases, the difference of variances between the IND structure and other structures becomes bigger. When the dependence among clusters is strong, we can improve the efficiency of our estimator by using appropriate working covariance structures. The variances under the AR1 are in general smaller than those from the EX structure, which is expected as the true covariance structure is AR1 in this simulation setting. Notice that the efficiency of our estimator is higher when the working covariance structure is closer to the truth as expected. Clearly, the proposed estimator is more efficient as interval-censoring rates are lower.

In the second simulation study, we assumed a clustered PIC failure time setting with identical regression coefficients across different clusters and identical marginal error distribution. We considered $n = 150$ clusters and each cluster size was fixed at two, i.e., $K = 2$. For cluster i, the failure time

Table 5: Summaries of results of semiparametric AFT models for (a) the original DRS dataset and (b) modified DRS dataset

Data	Effects	IND		EX		AR1	
		EST	SE	EST	SE	EST	SE
Original	Risk	−1.039	0.462	−1.005	0.462	−1.005	0.488
	Age	0.028	0.080	0.028	0.079	0.028	0.071
	Treatment	0.366	0.091	0.386	0.089	0.378	0.098
Modified	Risk	−1.191	0.512	−1.098	0.578	−1.095	0.552
	Age	0.057	0.080	0.065	0.072	0.068	0.072
	Treatment	0.415	0.197	0.394	0.163	0.394	0.149

$T_i = (T_{i1}, T_{i2})'$ was generated from

$$\log T_{ik} = X_{1ik} - X_{2ik} + \varepsilon_{ik},$$

where $X_{1ik} \sim N(0,1)$, X_{2ik} had three distinct numbers, 1, 2, 3, with probability 1/3 for each. The joint distribution of $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2})'$ was specified by a common marginal distribution. Only one marginal error distribution was considered: Normal distribution, $N(0,1)$. Other settings are equal to our first simulation study.

The simulation results are summarized in Table 4. Similar to the first simulation study, all estimators are unbiased. Furthermore, the sample standard error (SSE) is close to the average standard error (ASE). As the AR1 structure is the true covariance structure in our setting, variances under AR1 are smaller than those from other variance structures. For a given censoring percentage, as the dependence level increases, the difference in variances between IND and other variance structures becomes bigger. Especially, the difference in variances between IND and AR1 (the true variance structure) is big. This implies the fact that the efficiency of our estimator is higher when the working covariance structure is closer to the truth as expected.

4. An application to modified DRS data

The diabetic retinopathy study (DRS) started from 1971 with the aim to examine the effectiveness of laser treatment in delaying the onset of severe vision loss. Diabetic retinopathy is the most common and serious eye complication of diabetes, because this may result in poor vision or blindness. The 197 patients in this dataset were a 50% random sample of the patients with "high-risk" diabetic retinopathy, categorized by risk group 6 or higher is considered. Each patient had one eye randomized to laser treatment and the other eye received no treatment. For each eye, the event of interest was the time from initiation of treatment to the time when visual acuity dropped below 5/200 two visits in a row (defined as "blindness"). Other interests were the efficacy of the laser treatment and the influence of other risk factors. In addition to the treatment indicator, two covariates are available: age at diagnosis of diabetes and risk group (6 to 12, rescaled to 0.5 to 1.0). Right-censoring was caused by death, dropout, or the end of the study.

Originally, this dataset has exact and right-censored observations only. To illustrate our proposed method, we generated a clustered PIC version of this data. Specifically, we randomly chose observations among the exact data and we added and subtracted generated random numbers $U \sim \text{Uniform}(2,5)$ to randomly selected observations to make them interval-censored or left-censored. The numbers of left-censored, interval-censored, and right-censored observations are 45, 59, and 62, respectively. We fitted the AFT model to the original and modified DRS datasets. The results are summarized in Table 5. We reported estimated coefficients and their standard errors with IND, EX, and AR1 working weight matrix structures. The standard error of our estimator was based on 1000 bootstrap datasets. Comparing the two results, it can be seen that there is no significant difference in estimating the coefficient of effects under each working weight matrix structure, and standard error is also well estimated. Overall, the treatment remains statistically significant after data manipulation, increasing the log-failure time by about 0.4 years on average. The risk factor appears to be nearly significant, which means that the more risky group has a lower survival rate. Therefore, we can conclude that our method estimates the coefficients of the model without significant differences by approximating the censoring data very well, even though we made the censoring data from the exact data on purpose.

5. Discussion

In this article, we proposed a generalization of the Buckley-James method for imputing clustered interval-censored data and an iterative GEE procedure for the AFT model to accommodate potential correlation within clusters. Because of the merits of the GEE method, we do not need to estimate the exact working weight matrix Ω_i, and higher efficiency of our estimator may be achieved if Ω_i is properly chosen to be close enough to the true covariance of $\hat{Y}_i(b)$. Under regularity conditions, we may show that our estimator has asymptotic consistency and asymptotic normality. However, depending on the presence of exact observations, the theories can change substantially with possibly different convergence rates and thus are not pursued in this article. For real data applications, we modified the right-censored DRS data to create a PIC dataset. A possible example of clustered interval-censored failure time data can arise when some failure times of interest are correlated and clustered into groups due to sharing some common features, such as clinical sites, environmental factors, or certain unknown characteristics.

We can extend our method to other settings. Even if we consider our clusters have the same size K, we can assume each cluster has its own cluster size. For clustered failure times with unequal cluster sizes, the working weight matrix Ω_i has different dimensions for each cluster and can still be constructed with IND, EX, and AR1 structures. We can also consider different marginal error distributions. For example, in our simulation studies, we imposed the same marginal error distribution. Instead, the joint distribution of $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})'$ can be specified by three different marginal distributions. In this case, the magnitude of their correlation may not be well accessed and thus omitted in our studies.

In this article, we focused on the Buckley-James estimator, which is based on the least-square principle for imputed mean responses when the outcome is censored. Alternatively, one may consider rank-based inferences for interval-censored data (Choi and Choi, 2021; Choi *et al.*, 2023), which are generally more robust to model misspecification and potential outliers. Ritov (1990) showed that the Buckley-James estimator and the log-rank estimator are asymptotically equivalent for right-censored data. We expect that similar properties would hold for general interval-censored data. The benefit of the Buckley-James estimator is that the correlation structure can be incorporated into the model within the GEE framework as we showed in this article, which is generally not feasible with rank-based estimating methods. It would be interesting to further explore the relationship between two methods under interval-censoring.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (No. 2022M3J6A1063595, 2022R1A2C1008514).

References

Buckley J and James I (1979). Linear regression with censored data, *Biometrika*, **66**, 429–436.

Chen L, Sun J, and Xiong C (2016). A multiple imputation approach to the analysis of clustered interval-censored failure time data with the additive hazards model, *Computational Statistics & Data Analysis*, **103**, 242–249.

Chiou SH, Kang S, Kim J, and Yan J (2014). Marginal semiparametric multivariate accelerated failure time model with generalized estimating equations, *Lifetime Data Analysis*, **20**, 599–618.

Choi S and Huang X (2021). Efficient inferences for linear transformation models with doubly-censored data, *Communications in Statistics-Theory and Methods*, **50**, 2188–2200.

Choi T and Choi S (2021). A fast algorithm for the accelerated failure time model with high-dimensional time-to-event data, *Journal of Statistical Computation and Simulation*, **91**, 3385–3403.

Choi T, Choi S, and Bandyopadhyay D (2023). Rank-Based inferences for the accelerated failure time model with partially interval-censored data. Submitted.

Choi T, Kim AK, and Choi S (2021). Semiparametric least-squares regression with doubly-censored data, *Computational Statistics & Data Analysis*, **164**, 107306.

Efron B (1967). The two sample problem with censored data. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, 831–853.

Gao F and Chan KCG (2019). Semiparametric regression analysis of length-biased interval-censored data, *Biometrics*, **75**, 121–132.

Gao F, Zeng D, and Lin DY (2017). Semiparametric estimation of the accelerated failure time model with partly interval-censored data, *Biometrics*, **73**, 1161–1168.

Huang J (1996). Efficient estimation for the proportional hazards model with interval censoring, *The Annals of Statistics*, **24**, 540–568.

Huang J (1999). Asymptotic properties of nonparametric estimation based on partly interval-censored data, *Statistica Sinica*, **9**, 501–519.

Huang J and Wellner JA (1997). Interval-Censored survival data: A review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics*, Springer, New York.

Ji S, Peng L, Cheng Y, and Lai HC (2012). Quantile regression for doubly-censored data, *Biometrics*, **68**, 101–112.

Jin Z, Lin DY, and Ying Z (2006). On least-squares regression with censored data, *Biometrika*, **93**, 147–161.

Kalbfleisch JD and Prentice RL (2002). *The Statistical Analysis of Failure Time Data*, Wiley, New York.

Kim JS (2003). Maximum likelihood estimation for the proportional hazards model with partly interval-censored data, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **65**, 489–502.

Kim YJ, Cho H, Kim J, and Jhun M (2010). Median regression model with interval-censored data, *Biometrical Journal*, **52**, 201–208.

Kor CT, Cheng KF, and Chen YH (2013). A method for analyzing clustered interval-censored data based on cox's model, *Statistics in Medicine*, **32**, 822–832.

Lam KF, Xu Y, and Cheung TL (2010). A multiple imputation approach for clustered interval-censored survival data, *Statistics in Medicine*, **29**, 680–693.

Liang KY and Zeger SL (1986). Longitudinal data analysis using generalized linear models, *Biometri ka*, **73**, 13–22.

Lin G, He X, and Portnoy S (2012). Quantile regression with doubly-censored data, *Computational Statistics & Data Analysis*, **56**, 797–812.

Mao L, Lin DY, and Zeng D (2017). Semiparametric regression analysis of interval-censored competing risks data, *Biometrics*, **73**, 857–865.

McCulloch CE and Searle SR (2004). *Generalized, Linear, and Mixed Models*, John Wiley & Sons.

Peng L and Huang Y (2008). Survival analysis with quantile regression models, *Journal of the American Statistical Association*, **103**, 637–649.

Portnoy S (2003). Censored regression quantiles, *Journal of the American Statistical Association*, **98**, 1001–1012.

Ren JJ (2003). Regression m-estimators with non-iid doubly-censored data, *The Annals of Statistics*, **31**, 1186–1219.

Ren JJ and Gu M (1997). Regression m-estimators with doubly-censored data, *The Annals of Statistics*, **25**, 2638–2664.

Ritov Y (1990). Estimation in a linear regression model with censored data, *The Annals of Statistics*, **18**, 303–328.

Shen X (1998). Proportional odds regression and sieve maximum likelihood estimation, *Biometrika*, **85**, 165–177.

Sun J (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*, Springer, New Yrok.

Turnbull BW (1974). Nonparametric estimation of a survivorship function with doubly-censored data, *Journal of the American Statistical Association*, **69**, 169–173.

Turnbull BW (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **38**, 290–295.

Wellner JA 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.

Zeng D, Gao F, and Lin DY (2017). Maximum likelihood estimation for semiparametric regression models with multivariate interval-censored data, *Biometrika*, **104**, 505–525.

Zeng D, Mao L, and Lin D (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data, *Biometrika*, **103**, 253–271.

Zhao X, Zhao Q, Sun J, and Kim JS (2008). Generalized log-rank tests for partly interval-censored failure time data, *Biometrical Journal*, **50**, 375–385.

Received December 22, 2022; Revised February 22, 2023; Accepted March 10, 2023