HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2018 December 01.

Published in final edited form as:

Biometrics. 2017 December ; 73(4): 1161–1168. doi:10.1111/biom.12700.

Semiparametric Estimation of the Accelerated Failure Time Model with Partly Interval-Censored Data

Fei Gao[*], **Donglin Zeng**[**], and **Dan-Yu Lin**[***]

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, U.S.A

Summary

Partly interval-censored (PIC) data arise when some failure times are exactly observed while others are only known to lie within certain intervals. In this article, we consider efficient semiparametric estimation of the accelerated failure time (AFT) model with PIC data. We first generalize the Buckley–James estimator for right-censored data to PIC data. Then, we develop a one-step estimator by deriving and estimating the efficient score for the regression parameters. We show that under mild regularity conditions the generalized Buckley–James estimator is consistent and asymptotically normal and the one-step estimator is consistent and asymptotically normal with a covariance matrix that attains the semiparametric efficiency bound. We conduct extensive simulation studies to examine the performance of the proposed estimators in finite samples and apply our methods to data derived from an AIDS study.

Keywords

Bootstrap; Buckley–James estimator; Kernel estimation; One-step estimator; Semiparametric efficiency; Survival data

1. Introduction

Partly interval-censored (PIC) data consist of failure time observations, in which some of the failure times are exactly observed while others are only known to lie within certain intervals. Such data arise in clinical and epidemiological research when the occurrence of an asymptomatic event, such as diabetic nephropathy or HIV infection, is ascertained at clinic visits. If a subject takes frequent visits, then his or her failure time can be determined with sufficient accuracy. If the visits are infrequent, then the failure time is known to lie within an interval that may be too broad to be treated as exact.

Several statistical methods have been suggested to make inference with PIC data. Specifically, estimation of the survival function for PIC data was studied by Turnbull (1976) and Huang (1999), among others. Zhao et al. (2008) developed a generalized log-rank test

*fgao@live.unc.edu

**dzeng@email.unc.edu

***lin@bios.unc.edu

6. Supplementary Materials

The Web Appendices, referenced in Section 2, and the R package for the proposed estimators are available with this article at the *Biometrics* website on Wiley Online Library.

for PIC data and established its asymptotic properties. Kim (2003) studied nonparametric maximum likelihood estimation (NPMLE) for the proportional hazards model.

In this article, we consider the accelerated failure time (AFT) model, which relates the logarithm of the failure time linearly to the covariates (Kalbfleisch and Prentice, 1980, pp. 32–34). Because of its direct physical interpretation, the AFT model is an appealing alternative to the proportional hazards model, especially when the response variable does not pertain to failure time. It may provide a more accurate or more concise summarization of the data than the proportional hazards model in certain applications (Zeng and Lin, 2007). However, semiparametric estimation of the AFT model is highly challenging, even in the case of right-censored data (Prentice, 1978; Buckley and James, 1979; Tsiatis, 1990; Lai and Ying, 1991; Zeng and Lin, 2007; Lin and Chen, 2013). For PIC data, we first propose an iterative algorithm similar to that of Buckley and James (1979). We show that the resulting estimator is consistent and asymptotically normal and its variance can be consistently estimated by bootstrap. We then propose an efficient estimator for the (vector-valued) regression parameter by the one-step Newton–Raphson update with the efficient score. We derive the efficient score and construct the one-step estimator using kernel estimation. The one-step estimator is shown to be consistent and asymptotically normal, with a limiting covariance matrix that attains the semiparametric efficiency bound and can be consistently estimated through bootstrap. We conduct extensive simulation studies to examine the performance of the Buckley–James and one-step estimators in realistic settings, and we use our methods to analyze data derived from an AIDS clinical trial.

2. Methods

2.1. Data and Model

Let T denote the failure time and $\boldsymbol{X}$ denote a d-vector of covariates. The AFT model specifies that

$$\log T = \boldsymbol{X}^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon,$$

where $\boldsymbol{\beta}$ is a d-vector of unknown regression parameters, and ε is an unobserved error independent of $\boldsymbol{X}$. The distribution of ε is arbitrary such that the model is semiparametric.

Let Δ indicate, by the values 1 versus 0, whether T is observed exactly or not. For $\Delta = 0$, there is a sequence of examination times $0 < U_1 < U_2 < \ldots < U_K < \infty$ that gives rise to the interval (L,R), where $L = \max\{U_k : U_k \leq T; k = 0, \ldots, K\}$, and $R = \min\{U_k : U_k \geq T; k = 1, \ldots, K+1\}$, with $U_0 = 0$ and $U_{K+1} = \infty$. We assume that the proportion of $\Delta = 1$ is not negligible, and the joint distribution of $(U_1, \ldots, U_K)$ is independent of T given $\boldsymbol{X}$ and $\Delta = 0$. Note that $L = 0$ represents a left-censored observation and $R = \infty$ represents a right-censored observation. For a random sample of n subjects, the PIC data consist of

$$\{\Delta_i, \Delta_i T_i, (1-\Delta_i)L_i, (1-\Delta_i)R_i, \boldsymbol{X}_i\} \ (i = 1, \ldots, n).$$

2.2. Generalized Buckley–James Estimation

If the failure time is observed for every subject, then the classical least-squares estimator for $\boldsymbol{\beta}$ is the solution to the estimating equation

$$\sum_{i=1}^{n}\left(\boldsymbol{X}_i-\overline{\boldsymbol{X}}\right)\left\{\left(Y_i-\overline{Y}\right)-\left(\boldsymbol{X}_i-\overline{\boldsymbol{X}}\right)^{\mathrm{T}}\boldsymbol{\beta}\right\}=\mathbf{0}, \quad (1)$$

where $Y_i = \log T_i$, $\overline{Y}=n^{-1}\sum_{i=1}^{n}Y_i$, and $\overline{\boldsymbol{X}}=n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i$. In the presence of censoring, some values of Y_i are not observed. Following the approach of Buckley and James (1979), we replace the unobserved Y_i by the conditional mean given the observed data. The conditional mean $\hat{Y}_i(\boldsymbol{\beta}, F)$ is given by

$$\begin{aligned}
&\Delta_i Y_i+(1-\Delta_i)\,E\,(Y_i|\max\{U_{ik}:u_{ik}<T_i\}=L_i,\min\{U_{ik}:U_{ik}\geq T_i\}=R_i,\boldsymbol{X}_i,\Delta_i=0,L_i,R_i)\\
&=\Delta_i Y_i+(1-\Delta_i)\times\frac{E\,[E\,\{Y_i I\,(\max\{U_{ik}:U_{ik}<T_i\}=L_i,\min\{U_{ik}:U_{ik}\geq T_i\}=R_i)|U_1,\ldots,U_K,\boldsymbol{X}_i\}\,|\boldsymbol{X}_i]}{E\,\{\Pr\,(\max\{U_{ik}:U_{ik}<T_i\}=L_i,\min\{U_{ik}:U_{ik}\geq T_i\}=R_i|U_1,\ldots,U_K,\boldsymbol{X}_i)|\boldsymbol{X}_i\}}\\
&=\Delta_i Y_i+(1-\Delta_i)\times\frac{E\left[\sum_{k=1}^{K}E\,\{Y_i I(U_{ik}=L_i,U_{i,k+1}=R_i,L_i<T_i\leq R_i)|U_1,\ldots,U_K,\boldsymbol{X}_i,\Delta_i=0\}\,|\boldsymbol{X}_i,\Delta_i=0\right]}{E\left\{\sum_{k=1}^{K}\Pr\,(U_{ik}=L_i,U_{i,k+1}=R_i,L_i<T_i\leq R_i|U_1,\ldots,U_K,\boldsymbol{X}_i,\Delta_i=0)|\boldsymbol{X}_i,\Delta_i=0\right\}}\\
&=\Delta_i Y_i+(1-\Delta_i)\times\frac{E\,\{Y_i I(L_i<T_i\leq R_i)|\boldsymbol{X}_i,\Delta_i=0,L_i,R_i\}\;E\left[\sum_{k=1}^{K}I(U_{ik}=L_i,U_{i,k+1}=R_i)|\boldsymbol{X}_i,\Delta_i=0,L_i,R_i\right]}{\Pr(L_i<T_i\leq R_i|\boldsymbol{X}_i,\Delta_i=0,L_i,R_i)E\left[\sum_{k=1}^{K}I\,(U_{ik}=L_i,U_{i,k+1}=R_i)|\boldsymbol{X}_i,\Delta_i=0,L_i,R_i\right]}\\
&=\Delta_i Y_{\boldsymbol{\beta},i}+(1-\Delta_i)\frac{\int_{L_{\boldsymbol{\beta},i}}^{R_{\boldsymbol{\beta},i}}u dF(u)}{F(R_{\boldsymbol{\beta},i})-F(L_{\boldsymbol{\beta},i})}+\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta},
\end{aligned}$$

where $Y_{\boldsymbol{\beta},i}=Y_i-\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}$, $L_{\boldsymbol{\beta},i}=\log L_i-\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}$, $R_{\boldsymbol{\beta},i}=\log R_i-\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}$, and F is the distribution function of $\boldsymbol{\varepsilon}$. The third equality follows from the conditional independence of the failure time and the examination times. Replacement of Y_i in (1) by $\hat{Y}_i(\boldsymbol{\beta}, F)$

$$\sum_{i=1}^{n}\left(\boldsymbol{X}_i-\overline{\boldsymbol{X}}\right)\left[\left\{\hat{Y}_i(\boldsymbol{\beta},F)-\overline{Y}(\boldsymbol{\beta},F)\right\}-\left(\boldsymbol{X}_i-\overline{\boldsymbol{X}}\right)^{\mathrm{T}}\boldsymbol{\beta}\right]=\mathbf{0},$$

where $\overline{Y}(\boldsymbol{\beta},F)=n^{-1}\sum_{i=1}^{n}\hat{Y}_i(\boldsymbol{\beta},F)$. Because F is unknown, we replace F by the self-consistency estimator $\hat{F}_{\boldsymbol{\beta}}$ (Turnbull, 1976; Huang, 1999) based on the transformed PIC data $\{\Delta_i, \Delta_i Y_{\boldsymbol{\beta},i}, (1-\Delta_i)L_{\boldsymbol{\beta},i}, (1-\Delta_i)R_{\boldsymbol{\beta},i}\}$ ($i = 1, \ldots, n$). The estimator $\hat{F}_{\boldsymbol{\beta}}$ solves the self-consistency equation

$$\hat{F}_{\boldsymbol{\beta}}(t)=n^{-1}\sum_{i=1}^{n}\left\{\Delta_i I(Y_{\boldsymbol{\beta},i}\leq t)+(1-\Delta_i)\frac{\hat{F}_{\boldsymbol{\beta}}(R_{\boldsymbol{\beta},i}\wedge t)-\hat{F}_{\boldsymbol{\beta}}(L_{\boldsymbol{\beta},i}\wedge t)}{\hat{F}_{\boldsymbol{\beta}}(R_{\boldsymbol{\beta},i})-\hat{F}_{\boldsymbol{\beta}}(L_{\boldsymbol{\beta},i})}\right\}, \quad (2)$$

where $a \wedge b = \min(a, b)$. If all of the failure times are observed, the right-hand side of equation (2) is simply the empirical distribution function for $Y_{\boldsymbol{\beta}}$. When the failure times are subject to censoring, the right-hand side is the conditional probability of $Y_{\boldsymbol{\beta}} \leq t$ given the observed data under the probability measure induced by $\hat{F}_{\boldsymbol{\beta}}$. The generalized Buckley–James estimator $\hat{\boldsymbol{\beta}}$ is the root of $\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\beta}) = \mathbf{0}$, where

$$\boldsymbol{U}_n(\beta, b) = n^{-1} \sum_{i=1}^{n} \left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right) \left[\left\{ \hat{Y}_i(b, \hat{F}_b) - \overline{Y}(b, \hat{F}_b) \right\} - \left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)^{\mathrm{T}} \beta \right].$$

The function $\boldsymbol{U}_n(\boldsymbol{\beta}, \boldsymbol{\beta})$ is not continuous in $\boldsymbol{\beta}$, so it is difficult to directly solve the estimating equation. We propose an iterative algorithm. With $(\boldsymbol{\beta}^{(0)}, F^{(0)})$ as the starting value, the algorithm proceeds as follows:

1. at step m, solve the self-consistency equation (2) with $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m-1)}$ to obtain $F^{(m)} = \hat{F}_{\boldsymbol{\beta}^{(m-1)}}$;
2. update $\boldsymbol{\beta}$ with the equation $\boldsymbol{\beta}^{(m)} = \boldsymbol{L}_n(\boldsymbol{\beta}^{(m-1)}, F^{(m)})$, where

$$\boldsymbol{L}_n(\beta, F) = \left\{ \sum_{i=1}^{n} \left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)^{\otimes 2} \right\}^{-1} \times \left[\sum_{i=1}^{n} \left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right) \left\{ \hat{Y}_i(\beta, F) - \overline{Y}(\beta, F) \right\} \right]$$

   with $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}$; and
3. set $m = m + 1$, and repeat steps (a) and (b) until convergence.

Denote the resulting estimator of $(\boldsymbol{\beta}, F)$ as $(\hat{\boldsymbol{\beta}}, \hat{F})$, where $\hat{F} = \hat{F}_{\hat{\boldsymbol{\beta}}}$. In Web Appendix A, we show that $(\hat{\boldsymbol{\beta}}, \hat{F})$ is consistent for the true value $(\boldsymbol{\beta}_0, F_0)$ and asymptotically normal under mild regularity conditions. The covariance matrix for the limiting distribution is difficult to directly estimate due to the lack of an analytical form. Therefore, we approximate the asymptotic distribution by bootstrapping the observations $\{\Delta_i, \Delta_i T_i, (1 - \Delta_i) L_i, (1 - \Delta_i) R_i, \boldsymbol{X}_i\}$ $(i = 1, \ldots, n)$. Let $\hat{\boldsymbol{\beta}}^*$ be the generalized Buckley–James estimator of a bootstrap sample. In Web Appendix B, we show that the conditional distribution of $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$ given the data converges weakly to the asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta_0)$. The empirical distribution of $\hat{\boldsymbol{\beta}}^*$ can then be used to approximate the distribution of $\hat{\boldsymbol{\beta}}$. Confidence intervals for individual components of $\boldsymbol{\beta}_0$ can be constructed by the Wald method (with the variance of $\hat{\boldsymbol{\beta}}^*$) or from the empirical percentiles of $\hat{\boldsymbol{\beta}}^*$.

2.3. One-Step Efficient Estimation

We wish to develop an estimator for $\boldsymbol{\beta}$ that attains the semi-parametric efficiency bound for PIC data. Let $\tilde{\boldsymbol{l}}_{\boldsymbol{\beta}}(\mathcal{O}, \boldsymbol{\beta}, F)$ be the efficient score for $\boldsymbol{\beta}$ under the AFT model with the observed data $\mathcal{O} \equiv \{\Delta, \Delta T, (1 - \Delta) L, (1 - \Delta) R, \boldsymbol{X}\}$. We can construct a semiparametric efficient estimator through the one-step Newton–Raphson update (Bickel et al., 1993, pp. 40–45) of the generalized Buckley–James estimator $(\hat{\boldsymbol{\beta}}, \hat{F})$,

$$\tilde{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}+\left\{\mathbb{P}_n\tilde{\boldsymbol{l}}_\beta(\mathscr{O};\hat{\boldsymbol{\beta}},\hat{F})^{\otimes 2}\right\}^{-1}\left\{\mathbb{P}_n\tilde{\boldsymbol{l}}_\beta(\mathscr{O};\hat{\boldsymbol{\beta}},\hat{F})\right\}, \quad (3)$$

where $\mathbb{P}_n$ is the empirical measure.

According to the semiparametric efficiency theory (Bickel et al., 1993, chap. 3), the efficient score for $\boldsymbol{\beta}$ is the sum of the scores for $\boldsymbol{\beta}$ and F along the least favorable direction $\boldsymbol{g}$ that is orthogonal to the tangent set for F. After the derivations given in Web Appendix C, we find that the least favorable direction $\boldsymbol{g}$ satisfies an integral equation. We replace the unknown quantities in the integral equation by appropriate sample estimators. The resulting function $\boldsymbol{g}$ satisfies the linear equation $\boldsymbol{A}(\hat{\boldsymbol{g}}(t_1)^{\mathrm{T}},\ldots,\hat{\boldsymbol{g}}(t_m)^{\mathrm{T}})^{\mathrm{T}}=\boldsymbol{c}$, where $\boldsymbol{A}=(a_{jl})\in\mathbb{R}^{m\times m}$, $\boldsymbol{c}=(\boldsymbol{c}_1^{\mathrm{T}},\ldots,\boldsymbol{c}_m^{\mathrm{T}})^{\mathrm{T}}$,

$$a_{jl}=I(l=j)\hat{P}\left(\Delta|Y_{\boldsymbol{\beta}_0}=t_j\right)+\hat{F}\{t_l\}\mathbb{P}_n\times\left[(1-\Delta)\frac{I(L_{\hat{\boldsymbol{\beta}}}<t_j\le R_{\hat{\boldsymbol{\beta}}},L_{\hat{\boldsymbol{\beta}}}<t_l\le R_{\hat{\boldsymbol{\beta}}})}{\left\{\hat{F}(R_{\hat{\boldsymbol{\beta}}})-\hat{F}(L_{\hat{\boldsymbol{\beta}}})\right\}^2}\right],$$

$$\boldsymbol{c}_j=\hat{P}\left(\Delta\boldsymbol{X}|Y_{\boldsymbol{\beta}_0}=t_j\right)\frac{\hat{f}'(t_j)}{\hat{f}(t_j)}+\mathbb{P}_n\times\left[(1-\Delta)\frac{\boldsymbol{X}\left\{\hat{f}(R_{\hat{\boldsymbol{\beta}}})-\hat{f}(L_{\hat{\boldsymbol{\beta}}})\right\}I(L_{\hat{\boldsymbol{\beta}}}<t_j\le R_{\hat{\boldsymbol{\beta}}})}{\left\{\hat{F}(R_{\hat{\boldsymbol{\beta}}})-\hat{F}(L_{\beta})\right\}^2}\right],$$

$Y_{\boldsymbol{\beta}}=Y-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}$, $L_{\boldsymbol{\beta}}=\log L-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}$, $R_{\boldsymbol{\beta}}=\log R-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}$, and $\hat{F}\{t_l\}$ is the jump size of $\hat{F}$ at t_l. Let f_0 and f_0' be the density function of ε and its derivative, respectively. The terms $\hat{f}(t)$, $\hat{f}'(t)$, $\hat{P}(\Delta|Y_{\boldsymbol{\beta}_0}=t)$, and $\hat{P}(\Delta\boldsymbol{X}|Y_{\boldsymbol{\beta}_0}=t)$ are kernel estimators of f_0, f_0', $E(\Delta|Y_{\boldsymbol{\beta}_0}=t)$, and $E(\Delta\boldsymbol{X}|Y_{\boldsymbol{\beta}_0}=t)$, defined as

$$\hat{f}(t)=\frac{1}{a_n}\int_0^\infty K\left(\frac{s-t}{a_n}\right)d\hat{F}(s),$$

$$\hat{f}'(t)=\frac{1}{b_n^3\int u^2K(u)du}\int_0^\infty(s-t)K\left(\frac{s-t}{b_n}\right)d\hat{F}(s),$$

$$\hat{P}\left(\Delta|Y_{\boldsymbol{\beta}_0}=t\right)=\frac{1}{na_n\hat{f}(t)}\mathbb{P}_n\sum_{i=1}^n\Delta_iK\left(\frac{Y_{\hat{\boldsymbol{\beta}},i}-t}{a_n}\right),$$

and

$$\hat{P}\left(\Delta\boldsymbol{X}|Y_{\boldsymbol{\beta}_0}=t\right)=\frac{1}{na_n\hat{f}(t)}\sum_{i=1}^n\Delta_i\boldsymbol{X}_iK\left(\frac{Y_{\hat{\boldsymbol{\beta}},i}-t}{a_n}\right),$$

where $K(\cdot)$ is a smooth and symmetric kernel function, and a_n and b_n are bandwidths. The conditions for the choices of the kernel function and bandwidths can be found in Web Appendix D.

The efficient score function can be estimated by

$$\hat{l}(\hat{\boldsymbol{\beta}}, \hat{F}, \hat{\boldsymbol{g}}) = -\left[\Delta\left\{\boldsymbol{X}\frac{\hat{f}^{'}(Y_{\hat{\boldsymbol{\beta}}})}{\hat{f}(Y_{\hat{\boldsymbol{\beta}}})} + \hat{\boldsymbol{g}}(Y_{\hat{\boldsymbol{\beta}}})\right\} + (1-\Delta) \times \frac{\boldsymbol{X}\left\{\hat{f}(R_{\hat{\boldsymbol{\beta}}}) - \hat{f}(L_{\hat{\boldsymbol{\beta}}})\right\} + \int_{L_{\hat{\boldsymbol{\beta}}}}^{R_{\hat{\boldsymbol{\beta}}}} \hat{\boldsymbol{g}}(u) d\hat{F}(u)}{\hat{F}(R_{\hat{\boldsymbol{\beta}}}) - \hat{F}(L_{\hat{\boldsymbol{\beta}}})}\right].$$

We replace the efficient score function $\tilde{\boldsymbol{l}}_{\boldsymbol{\beta}}(\mathcal{O}; \hat{\boldsymbol{\beta}}, \hat{F})$ in (3) by $\hat{\boldsymbol{l}}(\hat{\boldsymbol{\beta}}, \hat{F}, \hat{\boldsymbol{g}})$ to obtain the one-step estimator

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \left\{\mathbb{P}_n \hat{\boldsymbol{l}}(\hat{\boldsymbol{\beta}}, \hat{F}, \hat{\boldsymbol{g}})^{\otimes 2}\right\}^{-1} \left\{\mathbb{P}_n \hat{\boldsymbol{l}}(\hat{\boldsymbol{\beta}}, \hat{F}, \hat{\boldsymbol{g}})\right\}.$$

In Web Appendix D, we show that $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ converges in distribution to a mean-zero normal random vector with a covariance matrix that attains the semiparametric efficiency bound. We estimate the covariance matrix by bootstrapping the observations and applying the one-step procedure. The validity of the bootstrap is proved in Web Appendix D. We also show that if the error $\boldsymbol{\varepsilon}$ is normally distributed, then the efficient score function is equivalent to the generalized Buckley–James estimating function. Thus, the generalized Buckley–James estimator is semiparametric efficient when the error is normally distributed.

3. Simulation Studies

We conducted extensive simulation studies to assess the performance of the proposed methods. We generated failure times from the AFT model: $\log T = -X_1 - X_2 - \varepsilon$, where X_1 and X_2 are independent Bernoulli(0.5) and standard normal variables, respectively, and ε is independent of (X_1, X_2). We considered four error distributions: standard normal distribution; standard extreme-value distribution; extreme-value distribution with location and scale parameters of −0.5 and 1.5, respectively; and logarithm of the gamma distribution with shape and scale parameters of 1 and 1, respectively. We simulated the time to loss to follow-up C from Uniform[10, 15]. For each subject, with probability p, we exactly observed the failure time T if $T \leq C$ and obtained a right-censored observation at C if $T > C$. With the remaining probability $1 - p$, we generated a sequence of examination times $U_k = U_{k-1} + \text{Uniform}[0.1, 1]$ $(k = 1, \ldots, K)$ such that $U_K < C$. We created the interval-censored observation $(L, R) \equiv (U_k, U_{k+1})$ if $U_k < T \leq U_{k+1}$ for $k = 0, \ldots, K$. The probability p depends on the covariates such that $p = p_0 - 0.1I(X_1 = 1)$, where p_0 was chosen to yield approximately 25 and 50% exact observations.

We considered the iterative algorithm convergent if both the norm of the difference for $\boldsymbol{\beta}$ and the integrated mean squared difference for F in two successive steps are less than 10^{-4} or the difference of the mean squared error $n^{-1}\sum_{i=1}^{n}\{\hat{Y}(\boldsymbol{\beta}, \hat{F}_{\boldsymbol{\beta}}) - \overline{Y}(\boldsymbol{\beta}, \hat{F}_{\boldsymbol{\beta}}) - \left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)^{\mathrm{T}}\boldsymbol{\beta}\}^2$ between two successive steps is less than 10^{-2}. In all the scenarios we considered, the non-convergence rate was less than 1%. We estimated the standard error using the Wald method based on 200 bootstrap data sets.

Table 1 summarizes the results of the generalized Buckley–James estimation for sample sizes $n = 250$ and 500. The bias of the parameter estimator is small and tends to decrease as n increases. The standard error estimator accurately reflects the true variation, and the confidence intervals have proper coverage probabilities.

With the generalized Buckley–James estimator as the initial estimator, we carried out the one-step efficient estimation, and the results are shown in Table 2. We chose the Gaussian kernel for convenience. The optimal bandwidths for estimating the density and its derivative are $a_n = (4/3)^{1/5}\sigma n^{-1/5}$ and $b_n = (4/5)^{1/7}\sigma n^{-1/7}$(Swanepoel, 1988), where σ is the sample standard deviation of $\{Y_{\hat{\boldsymbol{\beta}},i} : \Delta_i = 1\}$ ($i = 1, \ldots, n$). We replaced σ by the minimum of the sample standard deviation and the interquartile range divided by 1.34, as suggested by Silverman (1986, p. 48).

The one-step estimator tends to be slightly positively biased, and the bias gets smaller as n increases. In the case of the normal error distribution, the one-step estimator has slightly larger standard error than the generalized Buckley–James estimator. This is not surprising because both estimators are asymptotically efficient when the error distribution is normal and the one-step estimator involves kernel approximation of the least favorable direction. For other error distributions, the one-step estimator achieves up to 16% efficiency gain over the generalized Buckley–James estimator in terms of variance. The efficiency gain in terms of mean squared error of the estimators is similar. The standard error estimator becomes more accurate as n increases. The confidence intervals have satisfactory coverage probabilities.

PIC data often arise as an approximation to interval-censored data, where the observations with short intervals are treated as exactly observed failure times. We examined the performance of the proposed estimators in this practical setting. We simulated the failure time T and time to loss to follow-up C in the same manner as before. For each subject, we generated a sequence of examination times $U_k = U_{k-1} + \text{Uniform}[a, b]$ ($k = 1, \ldots, K$) such that $U_K < C$. We set $(a, b) = (0, 0.1)$ with probability p and $(a, b) = (0.1, 1)$ with probability $1 - p$. We created the interval-censored observation $(L,R) \equiv (U_k, U_{k+1})$ if $U_k < T \leq U_{k+1}$ for $k = 0, \ldots, K$. If the interval length $R - L$ is smaller than 0.1, we treated the observation as exactly observed failure time at the geometric mid-point $\sqrt{LR}$. In this case, $\Delta = I(R - L < 0.1)$, and the exact observations are approximations to the true failure times.

We display the results for the proposed estimators with 50% exact observations in Table 3. The generalized Buckley–James estimator and one-step estimator have reasonably small bias. The standard error estimators accurately reflect the true variation, and the confidence intervals have satisfactory coverage probabilities. The one-step estimator achieves up to 13% efficiency gain for some of the considered error distributions.

A naive approach to analyzing interval-censored data is to approximate all interval-censored observations by single values and then apply the methodology for potentially right-censored data. We examined this approach in the second simulation setting by treating each interval-censored observation as exact failure time at the right end or the mid-point of the interval and applying the original Buckley–James estimator. As shown in Table 4, both

approximations yield estimators with smaller standard error than the generalized Buckley–James and one-step estimators but induce severe bias in the parameter estimation.

4. An AIDS Example

We considered an AIDS Clinical Trial Group (ACTG) study (Goggins and Finkelstein, 2000). In this clinical trial, blood and urine samples were collected at clinical visits to test for the presence of opportunistic infection cytomegalovirus (CMV), which is also known as shedding of the virus. The blood and urine samples were originally scheduled to be collected about every 12 and 4 weeks, respectively. The CMV shedding times in both blood and urine are interval-censored in that the events are only known to occur between the last negative and first positive tests.

The data set consists of 204 HIV-infected patients with at least one blood and urine samples taken during the study. For CMV shedding time in blood, 7 patients have left-censored observations, 174 patients have right-censored observations, and 23 patients have interval-censored observations. For CMV shedding time in urine, the corresponding numbers are 49, 88, and 67. The data set also includes the patient's baseline CD4 cell count as an indicator of less than versus greater than 75 (cells/μl). It is of interest to determine whether the baseline CD4 cell count is predictive of CMV shedding time.

This data set was previously analyzed by Goggins and Finkelstein (2000) using the proportional hazards model for bivariate interval-censored data. To illustrate the proposed methods, we generated a PIC version of the data. Specifically, we defined the failure time as the minimum of the shedding times in blood and in urine. If the shedding times in blood and in urine are $(L_b, R_b]$ and $(L_u, R_u]$, then the failure time is known to lie within $(L_b \wedge L_u, R_b \wedge R_u]$. The numbers of left- , interval-, and right-censored observations are 51, 65, and 88, respectively. The interval lengths for the interval-censored observations range from 1 to 9 months. We treated interval-censored observations with interval lengths less than 2 months as exact observations at the geometric mid-point of the interval to obtain 46 exact observations.

We fit the AFT model to the generated PIC data. We estimated the standard error of the generalized Buckley–James estimator using the Wald method based on 1000 bootstrap data sets. We used the optimal bandwidths described in the previous section for the one-step estimation. For comparisons, we also fit the proportion hazards model using the NPMLE method described in Kim (2003). The results are summarized in Table 5.

The estimates of the regression parameter in the AFT model are negative and thus indicate that patients with higher CD4 cell counts tend to have longer time to CMV shedding. The one-step estimator yields a larger estimate of the effect size than the generalized Buckley–James estimator, with a slightly larger standard error estimate, resulting in a slightly smaller p-value. Not surprisingly, the estimate of the regression parameter under the proportional hazards model has an opposite sign.

5. Discussion

It is much more challenging, both computationally and theoretically, to deal with PIC data under the AFT model than under the proportional hazards model. We developed a generalization of the Buckley–James estimator and a one-step efficient estimator, both of which perform well in realistic settings. We tackled the theoretical challenges through careful use of modern empirical process theory and semiparametric efficiency theory.

A non-negligible proportion of exact observations is a crucial assumption for the proposed methods. It plays an important role in establishing the asymptotic properties. With this assumption, there are some subjects with exactly observed failure times, so the estimator for the survival function of ε can be estimated accurately at those points. This leads to the $\sqrt{n}$ convergence rate, a faster rate than with only interval-censored observations. Computationally, we let the survival function be a step function with jumps at the exact failure times. Without exact observations, a natural estimator for the survival function would be a step function with potential jumps at all interval endpoints, such that the likelihood becomes non-concave and the estimation becomes unstable.

In practice, certain bootstrap samples may contain too few or no exact observations. We suggest to delete those samples provided that they account for a small proportion of all bootstrap samples. An alternative strategy is to perform parametric bootstrap, which requires modeling of the censoring distribution (Efron and Tibshirani, 1993, pp. 90–92).

We used kernel estimation for density and its derivative in constructing the one-step estimator. The estimation for this one-dimensional distribution is relatively stable and accurate. If the density or its derivative is estimated with bias, the resulting function will depart from the efficient score function. However, the function is still a valid score function, such that the one-step estimator remains consistent.

For the accelerated failure time model with right-censored data, the rank-based estimator (Gehan, 1965), which solves the gradient of a weighted probability for the observed rank, can be easily calculated via the linear programming technique. Lin and Chen (2013) proposed a one-step efficient estimation procedure using the rank-based estimator as the initial estimator. For PIC data, due to the existence of interval-censored observations, we cannot recover the rank structure to obtain rank-based estimating equations.

In most medical studies, the events of interest are asymptomatic such that the failure times are intrinsically interval-censored. A common practice is to apply the methodology for right-censored data by treating the time of the first detection or the mid-point of the interval as the exact failure time. However, this strategy can induce severe bias in the estimation, as shown in our simulation studies. The PIC methodology as presented in this article provides a better approximation to interval-censored data by treating only the small intervals as exact observations.

It is extremely challenging to perform semiparametric regression analysis of interval-censored data without treating any observations as exact. Although progress has been made on the semiparametric analysis of interval-censored data under the AFT model (Rabinowitz,

Tsiatis, and Aragon, 1995; Murphy, van der Vaart, and Wellner, 1999; Shen, 2000; Betensky, Rabinowitz, and Tsiatis, 2001; Tian and Cai, 2006), efficient estimation has not been explored. Our proposed methods require a non-negligible proportion of exact observations, which is crucial in establishing the asymptotic properties and constructing the computation algorithm. Therefore, the approach cannot be trivially extended to the interval-censored data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Betensky RA, Rabinowitz D, Tsiatis AA. Computationally simple accelerated failure time regression for interval censored data. Biometrika. 2001; 88:703–711.

Bickel, PJ., Klaassen, CAJ., Ritov, Y., Wellner, JA. Efficient and Adaptive Estimation for Semiparametric Models. New York: Springer; 1993.

Buckley J, James I. Linear regression with censored data. Biometrika. 1979; 66:429–436.

Efron, B., Tibshirani, RJ. An Introduction to the Bootstrap. Boca Raton: Chapman & Hall/CRC; 1993.

Gehan EA. A generalized wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika. 1965; 52:203–223. [PubMed: 14341275]

Goggins WB, Finkelstein DM. A proportional hazards model for multivariate interval-censored failure time data. Biometrics. 2000; 56:940–943. [PubMed: 10985240]

Huang J. Asymptotic properties of nonparametric estimation based on partly interval-censored data. Statistica Sinica. 1999; 9:501–519.

Kalbfleisch, JD., Prentice, RL. The Statistical Analysis of Failure Time Data. New York: Wiley; 1980.

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

Lai TL, Ying Z. Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. The Annals of Statistics. 1991; 19:1370–1402.

Lin Y, Chen K. Efficient estimation of the censored linear regression model. Biometrika. 2013; 100:525–530.

Murphy SA, van der Vaart AW, Wellner JA. Current status regression. Mathematical Methods of Statistics. 1999; 8:407–425.

Prentice RL. Linear rank tests with right censored data. Biometrika. 1978; 65:167–179.

Rabinowitz D, Tsiatis A, Aragon J. Regression with interval-censored data. Biometrika. 1995; 82:501–513.

Shen X. Linear regression with current status data. Journal of the American Statistical Association. 2000; 95:842–852.

Silverman, BW. Density Estimation for Statistics and Data Analysis. London: Chapman & Hall/CRC; 1986.

Swanepoel JWH. Mean integrated squared error properties and optimal kernels when estimating a distribution function. Communications in Statistics–Theory and Methods. 1988; 17:3785–3799.

Tian L, Cai T. On the accelerated failure time model for current status and interval censored data. Biometrika. 2006; 93:329–342.

Tsiatis AA. Estimating regression parameters using linear rank tests for censored data. The Annals of Statistics. 1990; 18:354–372.

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

Zeng D, Lin DY. Efficient estimation for the accelerated failure time model. Journal of the American Statistical Association. 2007; 102:1387–1396.

Zhao X, Zhao Q, Sun J, Kim JS. Generalized log-rank tests for partly interval-censored failure time data. Biometrical Journal. 2008; 50:375–385. [PubMed: 18435504]

Table 1

Simulation results for the generalized Buckley–James estimator

			$n = 250$				$n = 500$			
Error distribution	**Exact rate (%)**		**Bias**	**SE**	**SEE**	**CP**	**Bias**	**SE**	**SEE**	**CP**
$N(0, 1)$	25	β_1	0.001	0.144	0.144	0.944	0.001	0.103	0.101	0.948
		β_2	−0.001	0.076	0.077	0.948	0.001	0.054	0.054	0.949
	50	β_1	0.000	0.136	0.135	0.948	0.001	0.094	0.095	0.946
		β_2	0.000	0.070	0.070	0.946	0.001	0.049	0.049	0.947
EV(0,1)	25	β_1	−0.006	0.172	0.170	0.952	−0.003	0.120	0.119	0.950
		β_2	−0.008	0.092	0.094	0.946	−0.002	0.066	0.066	0.945
	50	β_1	−0.002	0.167	0.166	0.949	−0.001	0.118	0.118	0.953
		β_2	−0.002	0.088	0.087	0.945	0.000	0.064	0.062	0.949
EV(−0.5,1.5)	25	β_1	−0.015	0.251	0.256	0.953	−0.005	0.180	0.178	0.949
		β_2	−0.013	0.139	0.140	0.947	−0.005	0.098	0.097	0.948
	50	β_1	0.003	0.250	0.249	0.950	0.000	0.176	0.176	0.948
		β_2	−0.003	0.130	0.130	0.945	−0.001	0.093	0.092	0.948
Gamma(1,1)	25	β_1	−0.007	0.174	0.170	0.948	−0.002	0.119	0.120	0.953
		β_2	−0.007	0.095	0.094	0.949	−0.003	0.066	0.066	0.948
	50	β_1	−0.002	0.165	0.166	0.950	0.001	0.116	0.118	0.945
		β_2	0.000	0.086	0.087	0.943	−0.001	0.062	0.062	0.945

Note: Bias and SE are the bias and standard error, respectively, of the parameter estimator; SEE is the mean of the standard error estimator; and CP is the coverage probability of the 95% confidence interval. EV(*a*,*b*) denotes the extreme-value distribution with location parameter *a* and scale parameter *b*. Gamma(*a*,*b*) denotes the logarithm of the gamma distribution with shape parameter *a* and scale parameter *b*. Each entry is based on 10,000 replicates.

Table 2

Simulation results for the one-step estimator

			$n = 250$					$n = 500$				
Error distribution	**Exact rate (%)**		**Bias**	**SE**	**SEE**	**CP**	**RE**	**Bias**	**SE**	**SEE**	**CP**	**RE**
$N(0, 1)$	25	β_1	0.012	0.146	0.146	0.947	0.978	0.008	0.104	0.102	0.948	0.954
		β_2	0.009	0.076	0.077	0.943	0.991	0.008	0.054	0.054	0.945	1.005
	50	β_1	0.009	0.138	0.137	0.944	0.962	0.007	0.095	0.096	0.946	1.000
		β_2	0.006	0.072	0.071	0.942	0.961	0.005	0.049	0.049	0.941	1.000
EV(0,1)	25	β_1	0.001	0.169	0.169	0.953	1.031	0.001	0.118	0.118	0.951	1.026
		β_2	0.000	0.089	0.092	0.947	1.089	0.002	0.064	0.064	0.945	1.050
	50	β_1	0.005	0.155	0.162	0.951	1.155	0.005	0.111	0.115	0.958	1.121
		β_2	0.003	0.084	0.085	0.949	1.090	0.004	0.060	0.061	0.954	1.068
EV(−0.5,1.5)	25	β_1	−0.002	0.247	0.255	0.953	1.036	0.005	0.178	0.176	0.953	1.001
		β_2	−0.001	0.136	0.137	0.948	1.045	0.003	0.095	0.095	0.947	1.037
	50	β_1	0.016	0.232	0.239	0.951	1.166	0.011	0.168	0.169	0.950	1.105
		β_2	0.007	0.123	0.125	0.944	1.129	0.007	0.088	0.089	0.949	1.094
Gamma(1,1)	25	β_1	0.000	0.169	0.169	0.951	1.053	0.001	0.118	0.118	0.951	1.022
		β_2	0.000	0.092	0.092	0.952	1.061	0.000	0.063	0.064	0.947	1.079
	50	β_1	0.005	0.158	0.161	0.956	1.087	0.006	0.112	0.116	0.953	1.110
		β_2	0.004	0.082	0.085	0.949	1.097	0.003	0.059	0.061	0.950	1.110

Note: Bias and SE are the bias and standard error, respectively, of the parameter estimator; SEE is the mean of the standard error estimator; CP is the coverage probability of the 95% confidence interval; and RE is the relative efficiency, defined as the ratio of the variance of the generalized Buckley–James estimator to that of the one-step estimator. EV(a,b) denotes the extreme-value distribution with location parameter a and scale parameter b. Gamma(a,b) denotes the logarithm of the gamma distribution with shape parameter a and scale parameter b. Each entry is based on 10,000 replicates.

Table 3

Simulation results for the PIC approximation

			Generalized Buckley–James				One-step				
Error distribution			**Bias**	**SE**	**SEE**	**CP**	**Bias**	**SE**	**SEE**	**CP**	**RE**
$N(0, 1)$	$n = 250$	β_1	−0.007	0.135	0.138	0.947	0.006	0.137	0.139	0.947	0.960
		β_2	−0.008	0.072	0.072	0.947	0.003	0.072	0.072	0.946	1.002
	$n = 500$	β_1	−0.007	0.098	0.097	0.952	0.004	0.099	0.098	0.952	0.978
		β_2	−0.007	0.051	0.051	0.946	0.001	0.051	0.051	0.949	0.994
EV(0,1)	$n = 250$	β_1	−0.006	0.163	0.161	0.949	0.008	0.154	0.155	0.953	1.126
		β_2	−0.005	0.085	0.084	0.949	0.006	0.081	0.081	0.949	1.118
	$n = 500$	β_1	−0.007	0.116	0.115	0.949	0.003	0.112	0.111	0.952	1.073
		β_2	−0.007	0.061	0.060	0.949	0.001	0.058	0.058	0.952	1.118
EV(−0.5,1.5)	$n = 250$	β_1	−0.006	0.233	0.231	0.949	0.014	0.223	0.224	0.949	1.089
		β_2	−0.005	0.122	0.121	0.950	0.014	0.117	0.117	0.948	1.086
	$n = 500$	β_1	0.005	0.164	0.166	0.948	0.018	0.158	0.161	0.948	1.078
		β_2	−0.004	0.085	0.087	0.944	0.009	0.083	0.084	0.948	1.071
Gamma(1,1)	$n = 250$	β_1	−0.004	0.163	0.161	0.947	0.005	0.153	0.155	0.948	1.136
		β_2	−0.006	0.084	0.085	0.944	0.006	0.079	0.081	0.948	1.122
	$n = 500$	β_1	−0.005	0.115	0.115	0.949	0.004	0.109	0.112	0.950	1.108
		β_2	−0.007	0.060	0.060	0.946	0.001	0.057	0.058	0.951	1.096

Note: Bias and SE are the bias and standard error, respectively, of the parameter estimator; SEE is the mean of the standard error estimator; CP is the coverage probability of the 95% confidence interval; and RE is the relative efficiency, defined as the ratio of the variance of the generalized Buckley–James estimator to that of the one-step estimator. EV(a,b) denotes the extreme-value distribution with location parameter a and scale parameter b. Gamma(a,b) denotes the logarithm of the gamma distribution with shape parameter a and scale parameter b. Each entry is based on 10,000 replicates.

Table 4

Simulation results for the original Buckley–James estimator

			Right end		Mid-point	
Error distribution			**Bias**	**SE**	**Bias**	**SE**
$N(0, 1)$	$n = 250$	β_1	0.291	0.115	0.131	0.124
		β_2	0.243	0.062	0.119	0.065
	$n = 500$	β_1	0.291	0.084	0.133	0.088
		β_2	0.243	0.044	0.119	0.046
EV(0,1)	$n = 250$	β_1	0.388	0.125	0.233	0.139
		β_2	0.319	0.069	0.195	0.074
	$n = 500$	β_1	0.390	0.090	0.233	0.099
		β_2	0.318	0.049	0.196	0.053
EV(−0.5,1.5)	$n = 250$	β_1	0.536	0.154	0.396	0.177
		β_2	0.432	0.084	0.322	0.093
	$n = 500$	β_1	0.537	0.111	0.401	0.124
		β_2	0.432	0.059	0.321	0.064
Gamma(1,1)	$n = 250$	β_1	0.390	0.127	0.233	0.141
		β_2	0.319	0.069	0.196	0.074
	$n = 500$	β_1	0.389	0.088	0.232	0.097
		β_2	0.319	0.048	0.196	0.052

Note: Bias and SE are the bias and standard error, respectively, of the parameter estimator. EV(*a*,*b*) denotes the extreme-value distribution with location parameter *a* and scale parameter *b*. Gamma(*a*,*b*) denotes the logarithm of the gamma distribution with shape parameter *a* and scale parameter *b*. Each entry is based on 10,000 replicates.

Table 5

Regression analysis for the ACTG study

Model	Est	Std error	*Z*-statistic	*p*-value	95% CI
Proportional hazards model	0.814	0.205	3.974	<0.0001	(0.412, 1.215)
AFT model					
Generalized Buckley–James	−1.212	0.335	−3.616	<0.0001	(−1.835, −0.560)
One-step	−1.256	0.343	−3.664	<0.0001	(−1.802, −0.563)

Note: 95% CI is the 95% confidence interval based on the Wald method (for proportional hazards model) or the empirical percentiles of the bootstrap samples (for AFT model).