Statistica Sinica **25** (2015), 863-877
doi:http://dx.doi.org/10.5705/ss.2013.313

LOCAL BUCKLEY-JAMES ESTIMATION FOR HETEROSCEDASTIC ACCELERATED FAILURE TIME MODEL

Lei Pang, Wenbin Lu and Huixia Judy Wang

Merck & Co., Inc, North Carolina State University and George Washington University

Abstract: In survival analysis, the accelerated failure time model is a useful alternative to the popular Cox proportional hazards model due to its easy interpretation. Current estimation methods for the accelerated failure time model mostly assume independent and identically distributed random errors, but in many applications the conditional variance of log survival times depend on covariates exhibiting some form of heteroscedasticity. In this paper, we develop a local Buckley-James estimator for the accelerated failure time model with heteroscedastic errors. We establish the consistency and asymptotic normality of the proposed estimator and propose a resampling approach for inference. Simulations demonstrate that the proposed method is flexible and leads to more efficient estimation when heteroscedasticity is present. The value of the proposed method is further assessed by the analysis of a breast cancer data set.

Key words and phrases: Accelerated failure time model, Buckley-James estimation, heteroscedasticity, kernel estimation, local Kaplan-Meier, Survival analysis.

1. Introduction

In survival analysis, the accelerated failure time model is an attractive alternative to the popular proportional hazards model for its simplicity and ease of interpretability. The conventional accelerated failure time model assumes a direct linear relationship between T_i, the survival time or some transformation thereof, and the covariates X_i:

$$T_i = \alpha + X_i^T \beta + \epsilon_i,\ i = 1, \cdots, n, \tag{1.1}$$

where α is the intercept, β is the p-dimensional vector of regression coefficients, and ϵ_i is the independent and identically distributed (*i.i.d.*) random error with mean zero. The accelerated failure time model is semiparametric as the distribution of ϵ_i is not specified.

A large number of estimation methods have been proposed for the accelerated failure time model. By assuming unconditional independence between survival and censoring times, the "synthetic data" approaches via the inverse probability of censoring weighted technique have been studied by Koul, Susarla, and Van Ryzin (1981), Leurgans (1987), and Fan and Gijbels (1994), among others. Many methods were developed based on the more relaxed assumption of conditional independence of survival and censoring times, including the Buckley-James estimator (Buckley and James (1979); Lai and Ying (1991); Zhou and Li (2008)), weighted rank estimators (Tsiatis (1990); Ritov (1990); Ying (1993); Zhou (2005)), and nonparametric maximum profile likelihood estimator (Zeng and Lin (2007)), to name a few. The asymptotic properties of these estimators and their associated inference procedures have been formally studied. In particular, Jin et al. (2003) and Jin, Lin, and Ying (2006) developed proper resampling procedures for variance estimation of the rank and Buckley-James estimators, respectively.

A key assumption in the accelerated failure time model is that the random errors are independently and identically distributed, and independent of the covariates, while, in many applications, the random errors depend on the covariates and exhibit some form of heteroscedasticity. Stare, Heinzl, and Harrell (2000) showed through simulations that the Buckley-James estimator is biased for the accelerated failure time model with heteroscedastic errors. Compared with the vast literature on the standard accelerated failure time model, much less work has been done for the accelerated failure time model with heteroscedastic errors. Chen and Khan (2000) studied estimation in censored regression models with nonparametric heteroscedasticity where censoring variables are fixed and observed. Zhou, Bathke, and Kim (2012) discussed an empirical likelihood inference approach for a heteroscedastic accelerated failure time model. Zhang and Davidian (2008) proposed an estimation approach for the accelerated failure time model via a flexible representation of the error distribution, and they discussed the extension to incorporate heteroscedasticity of a specific parametric form. Heuchenne and Van Keilegom (2007) developed an estimation procedure for censored polynomial regression, where synthetic data points were constructed in a way to accommodate heteroscedasticity. Liu and Lu (2009) proposed a weighted least squares method based on the synthetic data transformation (Fan and Gijbels (1994)), a linear combination of the transformations proposed in Koul, Susarla, and Van Ryzin (1981) and Leurgans (1987).

Motivated by the iterative least square representations of the Buckley-James estimator, we develop a local Buckley-James estimator for a heteroscedastic accelerated failure time model. The main idea is to recursively impute the censored survival times by estimating the conditional mean based on the local Kaplan-Meier estimate of the conditional survival functions. The regression coefficients are then estimated through ordinary least squares fit based on the imputed survival times. We establish the theoretical properties of the local Buckley-James estimator and propose a resampling procedure for inference. By allowing the heteroscedasticity to depend on the mean index, our proposed method is free of the curse of dimensionality.

The rest of the paper is organized as follows. Section 2 describes the heteroscedastic accelerated failure time model and the proposed local Buckley-James estimator. Section 3 presents the theoretical properties of the local Buckley-James estimator and introduces a resampling procedure for inference. In Section 4, we present a comprehensive simulation study to demonstrate the effectiveness of the local Buckley-James method. The proposed method is applied to a breast cancer data set in Section 5. Section 6 concludes the paper with a brief discussion. The theoretical proofs are given in a Web Appendix.

2. Proposed Method

2.1. The heteroscedastic accelerated failure time model

We consider the heteroscedastic accelerated failure time model

$$T_i = \alpha + X_i^T\beta + \sigma(X_i^T\beta)\epsilon_i,\ i = 1, \ldots, n, \tag{2.1}$$

where T_i is the survival time or some transformation thereof, and ϵ_i are *i.i.d.* random errors with mean zero and standard deviation one. The function $\sigma(X_i^T\beta)$ describes the error heteroscedasticity with $\sigma(\cdot)$ an unspecified nonparametric function.

One motivation for the model arises when the conditional variance of log-transformed survival times depends on the covariates through the fitted mean log-survival. Another occurs in the generalized linear model when the variance of the response variable is a function of the mean. Letting $\sigma = \sigma(X_i)$, an arbitrary function of the covariate vector X_i, is theoretically enticing, but this imposes technical difficulties since nonparametric estimation of the function $\sigma(X_i)$ or the conditional survival function given X_i is challenging for high dimensional X_i due to the curse of dimensionality. We focus here on the estimation of β in (2.1).

2.2. The local Buckley-James estimation

Due to right censoring, we only observe the triplets (Y_i, δ_i, X_i), where $Y_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$, and C_i is the corresponding censoring time. Throughout, we assume that T_i and C_i are conditionally independent given X_i.

When there is no censoring, the classical ordinary least squares (OLS) estimator $\hat{\beta}$ can be obtained by solving, for β,

$$n^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(T_i - X_i^T\beta) = 0, \tag{2.2}$$

where $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$. The intercept α can be estimated by $\hat{\alpha} = n^{-1}\sum_{i=1}^n e_i(\hat{\beta})$, where $e_i(b) = T_i - X_i^T b$ for a given vector b.

In the presence of censoring, the exact survival times T_i are unknown for censored cases with $\delta_i = 0$. Buckley and James (1979) proposed to replace the censored T_i in (2.2) with the estimate of $E(T_i|T_i \geq C_i, Y_i, X_i)$, the conditional expectation of T_i. Buckley-James estimator $\hat{\beta}$ can then be obtained by fitting least squares regression with the imputed survival times. However, the conventional Buckley-James estimator assumes homoscedastic errors and thus does not work for model (2.1).

To account for heteroscedasticity, we propose a local Buckley-James estimator. The method shares the same spirit as the Buckley-James estimator by imputing the censored survival time T_i by its estimated conditional mean. Under the heteroscedastic accelerated failure time model (2.1),

$$\begin{aligned} E(T_i|T_i \geq C_i, Y_i, X_i) &= E(e_i|T_i \geq C_i, Y_i, X_i^T\beta) + X_i^T\beta \\ &= \frac{\int_{Y_i - X_i^T\beta}^{\infty} u dF_\beta(u|X_i^T\beta)}{1 - F_\beta(Y_i - X_i^T\beta|X_i^T\beta)} + X_i^T\beta, \end{aligned} \tag{2.3}$$

where $F_\beta(u|v)$ is the unknown conditional cumulative distribution function of the residual $e_i \equiv e_i(\beta) = T_i - X_i^T\beta$ given $X_i^T\beta = v$: $F_\beta(u|v) = P(e_i \leq u|X_i^T\beta = v)$. In the presence of heteroscedastic errors, $F_\beta(u|X_i^T\beta)$ depends on $X_i^T\beta$ and cannot be estimated by the Kaplan-Meier estimate. We adopt the local Kaplan-Meier estimator (Dabrowska (1987)) to estimate $F_\beta(u|X_i^T\beta)$, which is then used to estimate β through iteration. The proposed local Buckley-James algorithm is as follows.

Step 1. Obtain an initial coefficient estimator $\hat{\beta}_0$, for instance, the Buckley-James estimator.

Step 2. At the ath iteration, impute the censored survival times T_i by

$$\tilde{Y}_i(\hat{\beta}_a) = \delta_i Y_i + (1-\delta_i)\hat{E}(T_i|T_i \geq C_i, Y_i, X_i^T\hat{\beta}_a), i = 1, 2, \ldots, n,$$

where

$$\hat{E}(T_i|T_i \geq C_i, Y_i, X_i^T\hat{\beta}_a) = X_i^T\hat{\beta}_a + \frac{\int_{\tilde{e}_i(\hat{\beta}_a)}^{\infty} u d\hat{F}_{\hat{\beta}_a}(u|X_i^T\hat{\beta}_a)}{1 - \hat{F}_{\hat{\beta}_a}\{\tilde{e}_i(\hat{\beta}_a)|X_i^T\hat{\beta}_a\}}, \tag{2.4}$$

with $\tilde{e}_i(b) = Y_i - X_i^T b$. In (2.4), $\hat{F}_b(u|X_i^T b)$ is the local Kaplan-Meier estimate of $F_b(u|X_i^T b)$, the conditional cumulative distribution function of the residual $e_i(b)$ given $X_i^T b$,

$$\hat{F}_b(t|X_i^T b) = 1 - \prod_{j:\tilde{e}_j(b)<t}^{n} \left\{1 - \frac{B_{nj}(X_i^T b)\delta_j}{\sum_{k=1}^n I\{\tilde{e}_k(b) \geq \tilde{e}_j(b)\} B_{nk}(X_i^T b)}\right\},$$

where $B_{nk}, k = 1, 2, \ldots, n$, is a sequence of non-negative weights with $\sum_{k=1}^{n} B_{nk} = 1$. We choose the Nadaraya-Watson type of weight for $B_{nk}(X_i^T b)$ (Nadaraya (1964)),

$$B_{nk}(X_i^T b) = \frac{K((X_i^T b - X_k^T b)/h_n)}{\sum_{l=1}^{n} K((X_i^T b - X_l^T b)/h_n)},$$

where h_n is the bandwidth such that $h_n \to 0$ as $n \to \infty$ and $K(\cdot)$ is a symmetric kernel function.

Step 3. Fit a least squares regression using the imputed survival times $\tilde{Y}_i(\hat{\beta}_a)$ to obtain the updated estimator

$$\hat{\beta}_{a+1} = \Big\{ \sum_{i=1}^{n} (X_i - \bar{X}_n)^{\otimes 2} \Big\}^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)\{\tilde{Y}_i(\hat{\beta}_a) - \bar{Y}_n(\hat{\beta}_a)\},$$

where $\bar{Y}_n(\hat{\beta}_a) = n^{-1} \sum_{i=1}^{n} \tilde{Y}_i(\hat{\beta}_a)$.

Step 4. Repeat Steps 2 and 3 until convergence is achieved. We denote the converged estimator as $\hat{\beta}_{LBJ}$.

Our convergence criteria is $\max_j |\hat{\beta}_{a+1,j} - \hat{\beta}_{a,j}| < 0.01$, where $\hat{\beta}_{a+1,j}$ and $\hat{\beta}_{a,j}$ are the jth components of $\hat{\beta}_{a+1}$ and $\hat{\beta}_a$, respectively. The number of iterations needed for convergence depends on such factors as sample size, number of covariates, and censoring pattern. Based on our numerical experience, the algorithm often converges within 10 iterations. One possible reason is that in Step 1, we use Buckley-James estimator as the initial estimator; it serves a good starting value although it is not consistent.

3. Asymptotic Properties and Inference

3.1. Asymptotic properties

The local Buckley-James estimator $\hat{\beta}_{LBJ}$ is the solution to

$$\begin{aligned} U_n(b) &\equiv \sum_{i=1}^{n} (X_i - \bar{X}_n)\{X_i^T b - \tilde{Y}_i(b)\} \\ &= \sum_{i=1}^{n} \Big\{ \int_{-\infty}^{\infty} t dY_i^x(t, b) + \int_{-\infty}^{\infty} \int_{t}^{\infty} \frac{1 - \hat{F}_{ib}(s)}{1 - \hat{F}_{ib}(t)} ds dJ_i^x(t, b) \Big\} = 0, \end{aligned}$$

where $Y_i^x(t, b) = (X_i - \bar{X}_n) I\{\tilde{e}_i(b) \geq t\}$ and $J_i^x(t, b) = (X_i - \bar{X}_n) I\{\tilde{e}_i(b) \geq t, \delta_i = 0\}$ are the indicator processes related to the ith observation, and $\hat{F}_{ib}(t)$ is the shorthand notation for $\hat{F}_b(t|X_i^T b)$. Since $U_n(b)$ is not a continuous function of b, to facilitate theoretical investigation, we take the local Buckley-James estimator $\hat{\beta}_{LBJ}$ as a zero-crossing of the estimating function $U_n(b)$. As in Lai and Ying (1991), we take $V_n(b)$ as a smooth approximation of $U_n(b)$,

$$V_n(b) = \sum_{i=1}^{n} \left\{ \int_{-\infty}^{\infty} tdEY_i^x(t,b) + \int_{-\infty}^{\infty} \int_{t}^{\infty} \frac{1 - F_{ib}(s)}{1 - F_{ib}(t)} dsdEJ_i^x(t,b) \right\},$$

where $F_{ib}(t)$ is defined after (A.1) in the Web Appendix as the limit of $\hat{F}_{ib}(t)$. We need some regularity conditions.

A1. $\sup_i \|X_i\| \leq M$, where M is a positive constant, and $\beta \in B_p(0,\rho)$, a p-dimensional ball in R^p centered at zero and with radius ρ. In addition, $X_i^T\beta$ has a differentiable and bounded density function $f_\mu(\cdot)$ and $\sigma(\cdot)$ is differentiable.

A2. For all v, $F_\beta(u|v)$ has a bounded twice-differentiable density $f_\beta(u|v)$. In addition, $\int_{-\infty}^{\infty} u^2 dF_\beta(u|v) < \infty$ and $\int_{-\infty}^{\infty} \{\dot{f}_\beta(u|v)\}^2 / f_\beta(u|v) du < \infty$, where $\dot{f}_\beta(u|v)$ is the first derivative of $f_\beta(u|v)$ with respect to u.

A3. The bandwidth satisfies $h_n = O(n^{-1/2+\kappa})$, where $0 < \kappa \leq 1/6$.

A4. The kernel function $K(\cdot)$ is Lipschitz continuous of order one and satisfies $\int K(u)du = 1$, $\int uK(u)du = 0$, $\int K^2(u)du < \infty$, and $\int u^2 K(u)du < \infty$.

A5. There exist some constants ν_1 and $\nu_2 > 0$ such that $P(C_i - X_i'\beta > \nu_1) = 0$ and $\inf_v F_\beta(\nu_1|v) > \nu_2$ for all v.

A6. For $0 < \lambda < 1/12$, $\lim_{n\to\infty} n^{-3/4} \{ \inf_{\|b\| \leq \rho, \|b-\beta\| \geq n^{-\lambda}} \|V_n(b)\| \} = \infty$.

A7. The first order derivative matrix $\mathbf{\Gamma}_n$ of $n^{-1}V_n(b)$ at β converges to a finite and nondegenerate matrix $\mathbf{\Gamma}$, as n goes to infinity.

Conditions A1$-$A5 are standard conditions. Condition A6 is assumed to ensure the consistency of the local Buckley-James estimator as the zero-crossing of $U_n(b)$ for $b \in B_p(0,\rho)$. Let $v(b)$ denote the limit of $n^{-1}V_n(b)$ as $n \to \infty$. As shown by Lai and Ying (1991), A6 is satisfied when $\mathbf{\Gamma}$ is nondegenerate, and $v(b) \neq 0$ for $b \neq \beta$ with $||b|| \leq \rho$. Condition A7 is needed to establish the asymptotic normality of the local Buckley-James estimator.

Theorem 1. *Under* A1$-$A7 *we have, as* $n \to \infty$*,*

(i) $\hat{\beta}_{LBJ} - \beta = o(n^{-\lambda})$ *a.s.*;

(ii) $\sqrt{n}(\hat{\beta}_{LBJ} - \beta) \xrightarrow{d} N(0, \mathbf{\Gamma}^{-1}\mathbf{\Sigma}(\mathbf{\Gamma}^{-1})')$, *where* $\mathbf{\Sigma}$ *is the asymptotic covariance matrix of* $n^{-1/2}U_n(\beta)$ *as defined in Lemma* 4 *of the Web Appendix.*

The proof of Theorem 1 is given in the Web Appendix available at `http://www.stat.sinica.edu.tw/statistica`.

3.2. Inference via resampling

Since the matrices $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ take complicated analytical forms and involve the unknown conditional error density function, it is impractical to directly estimate them to obtain the variance estimate of $\hat{\beta}_{LBJ}$. Jin, Lin, and Ying (2006) proposed a resampling procedure for estimating the variance of the regular Buckley-James estimator and showed its validity. We adopt a similar resampling approach for variance estimation of the local Buckley-James estimator, as follows.

We first generate positive random variables W_i, $i = 1, 2, \ldots, n$, with $E(W_i) = \mathrm{Var}\,(W_i) = 1$; these are used to introduce random perturbation into the local Buckley-James estimation. Thus, take

$$L^*(b) = \Big\{ \sum_{i=1}^{n} W_i (X_i - \bar{X_n})^{\otimes 2} \Big\}^{-1} \Big[\sum_{i=1}^{n} W_i (X_i - \bar{X_n}) \{ \tilde{Y}_i^*(b) - \bar{Y}_n^*(b) \} \Big],$$

where

$$\tilde{Y}_i^*(b) = \delta_i Y_i + (1 - \delta_i) \Big[\frac{\int_{\tilde{e}_i(b)}^{\infty} u d\hat{F}_b^*(u|X_i^T b)}{1 - \hat{F}_b^*\{\tilde{e}_i(b)|X_i^T b\}} + X_i^T b \Big],$$

$$\hat{F}_b^*(t|X_i^T b) = 1 - \prod_{j:\tilde{e}_j(b) \le t}^{n} \Big[1 - \frac{W_j B_{nj}(X_i^T b)\delta_j}{\sum_{k=1}^{n} W_k I\{\tilde{e}_k(b) \ge \tilde{e}_j(b)\} B_{nk}(X_i^T b)} \Big],$$

and $\bar{Y}_n^*(b) = n^{-1} \sum_{i=1}^{n} \tilde{Y}_i^*(b)$. Here $\hat{F}_b^*(t|X_i^T b)$ is a randomly perturbed version of the local Kaplan-Meier estimate, $\tilde{Y}_i^*(b)$ is the imputed survival time based on $\hat{F}_b^*(t|X_i^T b)$, and $L^*(b)$ is the result of a perturbed least squares estimation.

Starting from the initial estimate $\hat{\beta}_{LBJ}$, $\hat{\beta}^*$ is obtained by iteratively updating $L^*(\cdot)$ from the previous estimate till the algorithm converges. As in Jin, Lin, and Ying (2006), it can be shown that given the observed data, $\sqrt{n}(\hat{\beta}^* - \hat{\beta}_{LBJ})$ converges to the same limiting distribution as $\sqrt{n}(\hat{\beta}_{LBJ} - \beta)$. The proof of this follows the proof of Theorem 1, and thus is omitted in the paper. By repeating the above resampling scheme N times, we obtain the resampled estimates $\hat{\beta}_k^*, k = 1, 2, \ldots, N$. The sample variance of $\{\hat{\beta}_k^*, k = 1, 2, \ldots, N\}$ provides a consistent variance estimate of $\hat{\beta}_{LBJ}$.

4. Simulation Study

In the first simulation study, we compare the finite sample performance of the Buckley-James (BJ) estimator, the Weighted Least Squares (WLS) estimator of Liu and Lu (2009), and our proposed local Buckley-James (LBJ) estimator in various situations. In Scenario 1, data are generated with homoscedastic errors and covariate-independent censoring. In Scenario 2, data are generated with heteroscedastic errors with covariate-independent censoring. Scenario 3 is based on the simulation setting in Liu and Lu (2009), which assumes heteroscedastic error and covariate-independent censoring but with more covariates. In Scenario 4, data are generated with heteroscedastic errors with covariate-dependent censoring.

For Scenarios 1, 2, and 4, we generated T_i, the log survival time, from the model

$$T_i = X_i^T\beta + \sigma(X_i^T\beta)\epsilon_i,\ i = 1, \ldots, n, \tag{4.1}$$

where $X_i = (X_{i1}, X_{i2})^T$, $X_{i1} \sim$ Unif(-1,1), and $X_{i2} \sim$ Bernoulli(0.5). Here we chose $\beta = (\beta_1, \beta_2)^T = (1, 1)^T$ and $\sigma(X_i^T\beta) = \exp(-0.3 - X_i^T\beta)$ for Scenarios 2 and 4, while $\sigma(X_i^T\beta) = 0.7$ for Scenario 1. We considered two different families of error distribution for ϵ_i: standard normal and centered standard extreme distributions. For Scenario 3, T_i was generated from (4.1) with $X_i = (1, X_{i1}, X_{i2}, X_{i3}, X_{i4})^T$, where $X_{i1} \sim$ Unif(-1,1), $X_{i2} = X_{i1}/3 + 2X_{i5}/3$ with $X_{i5} \sim$ Triangle(-2,2) being independent of X_{i1}, and X_{i3} and X_{i4} independent Bernoulli random variables with success probability 0.5, both independent of X_{i1} and X_{i2}. Here $\beta = (6, -1, 2, 1, -1)^T$, $\sigma(X_i^T\beta) = \exp(3.52 - X_i^T\beta)$, and $\epsilon_i \sim \mathrm{N}(0, 1)$.

For Scenarios 1, 2, and 3, the censoring time C_i was $\mathrm{N}(c_1, 2)$. For Scenario 4, the censoring time C_i was $\mathrm{N}(c_2, 2)$ if $X_{i2} = 1$, and $\mathrm{N}(c_3, 2)$ if $X_{i2} = 0$. Constants $c_i, i = 1, \cdots, 3$ were chosen to yield censoring proportions 20% and 40%: $c_1 = 2.4$, $c_2 = 1.6$, and $c_3 = 2.9$ for 20% censoring; $c_1 = 1.1$, $c_2 = 0.6$, and $c_3 = 1.9$ for 40% censoring. For each setting, we considered two sample sizes: $n = 200$ and 400. For the proposed LBJ method, the bandwidth parameter was $h_n = 4\mathrm{sd}(X^T\hat{\beta}_0)n^{-1/3}$, where $\mathrm{sd}(X^T\hat{\beta}_0)$ is the standard deviation of the linear index $X^T\hat{\beta}_0$, and $\hat{\beta}_0$ refers to the initial estimator in the LBJ algorithm. The variance of the LBJ estimator was computed based on 500 resamplings using the proposed resampling method with perturbation variables generated from the standard exponential distribution.

Tables 1−4 summarize the simulation results of three different methods in the four scenarios. In the tables, *bias* is the mean bias averaged over 500 simulations, *sd* is the Monte Carlo standard deviation, *se* is the mean estimated standard error obtained from the resampling procedure, and *covp* is the empirical coverage probability of the Wald-type 95% confidence interval. In Scenario 1, with homoscedastic error and covariate independent censoring, all three methods give essentially unbiased estimation. The resampling procedure for the LBJ method works reasonably well. The resampling standard errors are close to the Monte Carlo standard deviations, and the confidence intervals have coverage probabilities close to the 95% nominal level. Not surprisingly, BJ is slightly more efficient than LBJ as the *i.i.d.* error assumption is satisfied in this scenario. Both BJ and LBJ estimators tend to be more efficient than WLS. One possible explanation is that the imputation procedure in LBJ utilizes the information from the censored data more efficiently than WLS, which is partly dependent on the inverse probability weighting principle. In Scenario 2, with heteroscedastic errors, the

Table 1. Simulation Scenario 1: Homoscedastic Error and Covariate Independent Censoring.

			LBJ				BJ		WLS	
n	CP	Coef	*bias*	*sd*	*se*	*covp*	*bias*	*sd*	*bias*	*sd*
					Normal Error					
200	20%	β_1	0.004	0.095	0.096	0.958	0.002	0.092	0.001	0.149
		β_2	0.002	0.110	0.112	0.934	-0.001	0.108	0.013	0.159
	40%	β_1	0.002	0.102	0.107	0.948	0.002	0.101	-0.012	0.232
		β_2	-0.003	0.122	0.124	0.944	-0.002	0.120	0.018	0.274
400	20%	β_1	0.005	0.068	0.070	0.932	0.004	0.068	0.005	0.101
		β_2	0.006	0.078	0.082	0.942	0.004	0.078	0.003	0.109
	40%	β_1	0.007	0.079	0.079	0.948	0.005	0.076	-0.003	0.168
		β_2	0.005	0.094	0.091	0.948	0.004	0.088	-0.005	0.184
					Extreme Error					
200	20%	β_1	0.006	0.094	0.094	0.954	0.006	0.096	0.009	0.148
		β_2	-0.003	0.112	0.109	0.942	-0.003	0.114	0.004	0.158
	40%	β_1	0.006	0.107	0.106	0.948	0.004	0.109	0.012	0.245
		β_2	-0.003	0.130	0.123	0.944	-0.005	0.132	0.011	0.269
400	20%	β_1	-0.003	0.066	0.066	0.952	-0.003	0.067	-0.004	0.096
		β_2	-0.002	0.076	0.077	0.954	-0.002	0.077	-0.005	0.109
	40%	β_1	-0.004	0.077	0.075	0.930	-0.005	0.081	-0.012	0.164
		β_2	-0.007	0.088	0.087	0.926	-0.007	0.092	-0.014	0.182

BJ: Buckley-James estimator; LBJ: local Buckley-James estimator; WLS: weighted least squares estimator; *bias*: mean bias averaged over 500 simulations; *sd*: the Monte Carlo standard deviation; *se*: the mean estimated standard error obtained from the resampling procedure; *covp*: the coverage probability of the resampling 95% confidence interval.

Table 2. Simulation Scenario 2: Heteroscedastic Error and Covariate Independent Censoring.

			LBJ				BJ		WLS	
n	CP	Coef	*bias*	*sd*	*se*	*covp*	*bias*	*sd*	*bias*	*sd*
					Normal Error					
200	20%	β_1	0.015	0.117	0.124	0.928	0.066	0.119	0.019	0.162
		β_2	0.009	0.116	0.122	0.936	0.059	0.118	0.032	0.167
	40%	β_1	0.034	0.126	0.134	0.924	0.137	0.137	0.016	0.255
		β_2	0.025	0.130	0.133	0.920	0.126	0.138	0.044	0.288
400	20%	β_1	0.015	0.084	0.091	0.946	0.065	0.088	0.015	0.111
		β_2	0.012	0.082	0.089	0.954	0.061	0.086	0.012	0.112
	40%	β_1	0.026	0.090	0.099	0.940	0.129	0.099	0.014	0.184
		β_2	0.019	0.087	0.097	0.946	0.120	0.097	0.009	0.190
					Extreme Error					
200	20%	β_1	0.012	0.121	0.124	0.936	0.057	0.128	0.017	0.158
		β_2	0.003	0.119	0.121	0.924	0.046	0.127	0.015	0.162
	40%	β_1	0.024	0.129	0.131	0.946	0.124	0.150	0.018	0.263
		β_2	0.012	0.129	0.130	0.948	0.110	0.147	0.022	0.279
400	20%	β_1	0.001	0.085	0.088	0.944	0.044	0.091	-0.002	0.105
		β_2	-0.001	0.082	0.085	0.944	0.042	0.088	-0.003	0.116
	40%	β_1	0.008	0.092	0.094	0.940	0.106	0.109	-0.008	0.177
		β_2	0.002	0.089	0.091	0.950	0.100	0.106	-0.010	0.190

BJ: Buckley-James estimator; LBJ: local Buckley-James estimator; WLS: weighted least squares estimator; *bias*: mean bias averaged over 500 simulations; *sd*: the Monte Carlo standard deviation; *se*: the mean estimated standard error obtained from the resampling procedure; *covp*: the coverage probability of the resampling 95% confidence interval.

Table 3. Simulation Scenario 3: Heteroscedastic Error and Covariate-Independent Censoring with 4 covariates.

	LBJ				BJ		WLS		LS	
	bias	*sd*	*se*	*covp*	*bias*	*sd*	*bias*	*sd*	*bias*	*sd*
					Normal Error					
					n=200; cp=20%					
β_1	-0.007	0.136	0.112	0.944	-0.069	0.148	-0.003	0.113	0.004	0.168
β_2	0.006	0.226	0.199	0.936	0.120	0.254	0.012	0.204	-0.003	0.280
β_3	0.009	0.162	0.151	0.934	0.069	0.167	0.017	0.204	0.006	0.262
β_4	0.002	0.154	0.145	0.950	-0.082	0.171	-0.023	0.199	0.003	0.241
					n=200; cp=40%					
β_1	-0.016	0.150	0.127	0.938	-0.135	0.189	-0.005	0.215	0.010	0.251
β_2	0.025	0.245	0.222	0.928	0.243	0.331	0.018	0.381	-0.014	0.431
β_3	0.027	0.191	0.179	0.930	0.142	0.217	0.012	0.346	0.003	0.411
β_4	0.002	0.155	0.166	0.950	-0.170	0.222	-0.024	0.385	0.008	0.420
					n=400; cp=20%					
β_1	-0.009	0.093	0.094	0.958	-0.060	0.106	-0.009	0.079	-0.007	0.113
β_2	0.005	0.173	0.163	0.956	0.105	0.194	0.010	0.145	0.004	0.202
β_3	0.006	0.117	0.120	0.962	0.060	0.122	0.007	0.135	0.006	0.178
β_4	0.002	0.108	0.115	0.960	-0.072	0.128	-0.015	0.134	-0.008	0.165
					n=400; cp=40%					
β_1	-0.017	0.108	0.105	0.960	-0.113	0.139	-0.016	0.155	-0.009	0.165
β_2	0.017	0.194	0.180	0.948	0.205	0.255	0.003	0.260	-0.011	0.285
β_3	0.017	0.140	0.141	0.950	0.117	0.156	0.006	0.247	0.002	0.292
β_4	0.002	0.117	0.127	0.940	-0.143	0.172	-0.019	0.258	-0.013	0.289

BJ: Buckley-James estimator; LBJ: local Buckley-James estimator; WLS: weighted least squares estimator; LS: least squares estimator; *bias*: mean bias averaged over 500 simulations; *sd*: the Monte Carlo standard deviation; *se*: the mean estimated standard error obtained from the resampling procedure; *covp*: the coverage probability of the resampling 95% confidence interval.

Table 4. Simulation Scenario 4: Heteroscedastic Error and Covariate Dependent Censoring.

			LBJ				BJ		WLS	
n	CP	Coef	*bias*	*sd*	*se*	*covp*	*bias*	*sd*	*bias*	*sd*
						Normal Error				
200	20%	β_1	0.016	0.117	0.124	0.926	0.076	0.118	0.055	0.148
		β_2	0.000	0.123	0.129	0.938	0.051	0.126	0.331	0.180
	40%	β_1	0.036	0.124	0.128	0.926	0.145	0.132	0.051	0.252
		β_2	0.020	0.128	0.134	0.936	0.115	0.137	0.690	0.286
400	20%	β_1	0.009	0.085	0.096	0.948	0.068	0.081	0.043	0.109
		β_2	0.001	0.082	0.096	0.972	0.048	0.080	0.338	0.120
	40%	β_1	0.024	0.088	0.101	0.940	0.135	0.091	0.030	0.181
		β_2	0.013	0.088	0.105	0.956	0.109	0.093	0.695	0.188
						Extreme Error				
200	20%	β_1	0.011	0.114	0.121	0.966	0.063	0.120	0.056	0.153
		β_2	0.003	0.113	0.120	0.962	0.045	0.120	0.340	0.181
	40%	β_1	0.027	0.125	0.128	0.946	0.138	0.145	0.057	0.248
		β_2	0.010	0.124	0.129	0.954	0.101	0.142	0.692	0.281
400	20%	β_1	0.005	0.076	0.087	0.956	0.058	0.081	0.039	0.106
		β_2	-0.002	0.078	0.085	0.954	0.037	0.084	0.333	0.127
	40%	β_1	0.013	0.083	0.094	0.964	0.124	0.099	0.019	0.172
		β_2	0.000	0.085	0.093	0.946	0.090	0.101	0.682	0.200

BJ: Buckley-James estimator; LBJ: local Buckley-James estimator; WLS: weighted least squares estimator; *bias*: mean bias averaged over 500 simulations; *sd*: the Monte Carlo standard deviation; *se*: the mean estimated standard error obtained from the resampling procedure; *covp*: the coverage probability of the resampling 95% confidence interval.

estimates from BJ are clearly biased, and the bias is more prominent with heavier censoring. In contrast, both LBJ and WLS give unbiased estimations. As in Scenario 1, LBJ tends to be more efficient than WLS. The design of Scenario 3 is the same as the example used in Liu and Lu (2009). For comparison, in Table 3 we also include the results of the unweighted version (LS) of the weighted least squares method. Compared with LS, WLS accounts for the heteroscedasticity by performing a weighted least squares based on the nonparametric estimates of the error conditional variance, and leads to more efficient estimation. For the light censoring (20%), WLS and LBJ estimates have similar performances. For heavier censoring (40%), LBJ estimates have smaller variances than WLS, which agrees with our observations in Scenarios 1-2. Scenario 4 presents a more complicated design, where the errors are heteroscedastic and the censoring depends on the covariates, violating assumptions required by BJ and WLS. As observed in Scenarios 2-3, BJ estimates show systematic biases. The WLS estimates are also biased, especially for the estimation of β_2, the coefficient of the covariate that affects the censoring distribution.

We conducted a second simulation study to compare our proposed method with that of Heuchenne and Van Keilegom (2007), denoted by HV. We considered the same simulation setting for the normal heteroscedastic regression model with one predictor as studied in Table 3 for model (12) of Heuchenne and Van Keilegom (2007). The simulation results are summarized in Table 5, with results for the BJ and HV estimators copied from Table 3 of Heuchenne and Van Keilegom (2007). We see that here the LBJ and HV methods perform similarly in terms of efficiency. The method of Heuchenne and Van Keilegom (2007) does rely on the nonparametric estimation of the conditional distribution of T_i given a univariate covariate, and is relatively difficult to generalize to multiple covariates due to the curse of dimensionality in nonparametric estimation.

The proposed LBJ method works competitively well in our simulations, for homoscedastic or heteroscedastic error, and covariate-dependent or covariate-independent censoring. It tends to be more efficient than WLS, possibly due to the efficient imputation step of the LBJ method.

5. Breast Cancer Data Analysis

To illustrate our method, we analyzed a breast cancer data set. The breast cancer data is from a clinical trial with three treatment arms of adjunct therapies for breast cancer (Farewell (1986)). Besides the observed survival times and censoring indicator, the data set also contains indicators for two treatments, trt1 and trt2 (control arm is set as the baseline), clinical stage I indicator (an early stage indicator, equals 1 with tumor size smaller than 2 cm and no positive movable axillary nodes), and the number of lymph nodes having disease involvement.

Table 5. Comparison with Heuchenne and Van Keilegom's method.

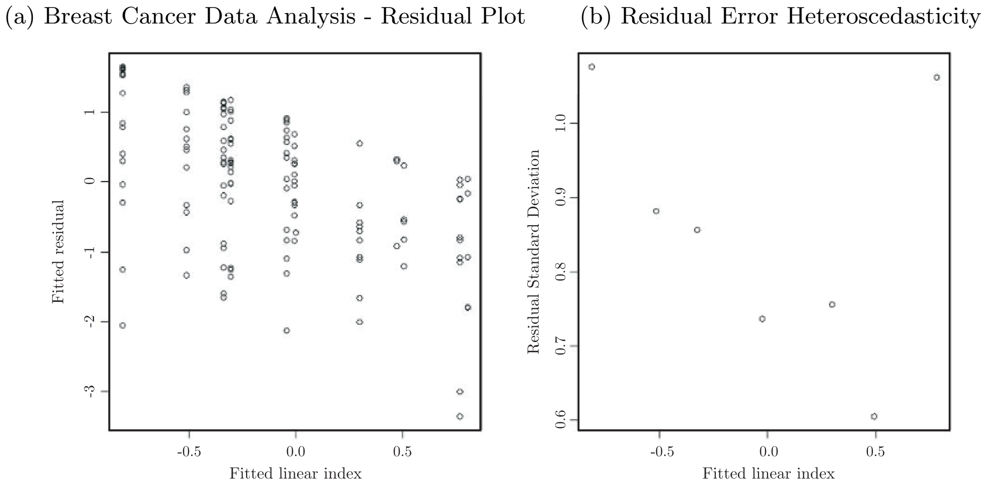
	β_0			β_1		
	Bias	Var	MSE	Bias	Var	MSE
	α_0=0.7, α_1=9.85, ρ=1, γ=1					
BJ	0.085	0.007	0.014	0.181	0.049	0.082
HV	0.063	0.006	0.010	0.135	0.048	0.066
LBJ	0.035	0.007	0.009	-0.077	0.063	0.069
	α_0=1.5, α_1=9.5, ρ=2, γ=2					
BJ	0.166	0.027	0.054	0.366	0.197	0.331
HV	0.100	0.023	0.033	-0.266	0.190	0.260
LBJ	0.055	0.028	0.031	-0.141	0.231	0.251
	α_0=2.4, α_1=10, ρ=4, γ=3					
BJ	0.230	0.061	0.114	0.475	0.466	0.692
HV	0.119	0.052	0.066	-0.326	0.444	0.550
LBJ	0.088	0.055	0.063	-0.154	0.482	0.505
	α_0=2.6, α_1=10, ρ=4, γ=5					
BJ	0.465	0.172	0.388	0.967	1.200	2.130
HV	0.201	0.136	0.177	-0.569	1.160	1.480
LBJ	0.198	0.168	0.207	-0.512	1.430	1.692

BJ: Buckley-James estimator; LBJ: local Buckley-James estimator; HV: Heuchenne and Van Keilegom's estimator; Bias: mean bias averaged over 500 simulations; Var: the Monte Carlo variance of 500 estimates; MSE: the mean squared error of 500 estimates.

The data set contains 139 records with 44 events (censoring proportion is about 68%). To simplify the analysis, we transformed the number of lymph nodes to a binary variable corresponding to lymph nodes having disease involvement.

As a preliminary analysis, we used the BJ method to fit the data and obtain the estimator $\hat{\beta}_{BJ}$. With $Y_i - X_i^T\hat{\beta}_{BJ}$ as the estimated residual and $X_i^T\hat{\beta}_{BJ}$ as the estimated linear index, Y_i the observed log survival time, we plotted the centered residuals against the estimated linear indices, as shown in panel (a) of Figure 1. Although the residual plot does not truthfully describe the error heteroscedasticity since it plots censored residuals together with uncensored ones, it suggests the existence of error heteroscedasticity. Based on the plot, we can see that the estimated linear indices mainly focus around 7 values. To better illustrate the error heteroscedasticity, we plotted the standard deviations of the fitted residuals around these values. The plot is given in panel (b) of Figure 1, and shows that the residual variance is not a constant but rather depends on the linear index in some nonlinear fashion.

We applied our method to the breast cancer data, and compared it with the BJ and WLS methods. The results from methods BJ, LBJ, and WLS are summarized in Table 6. Methods BJ and LBJ give similar coefficient estimates. For predictors trt2 and clinical stage I, the WLS estimates are different from those of BJ and LBJ. In general, the LBJ estimates have smaller standard errors than



(a) Breast Cancer Data Analysis - Residual Plot

(b) Residual Error Heteroscedasticity

Figure 1. Residual and error heteroscedasticity plots for breast cancer data.

Table 6. Data analysis results: point estimation, standard error and p-value.

	LBJ			BJ			WLS		
	coeff	se	p-val	coeff	se	p-val	coeff	se	p-val
trt1	0.506	0.192	0.008	0.480	0.224	0.032	0.416	0.223	0.062
trt2	0.471	0.184	0.010	0.449	0.207	0.030	0.277	0.200	0.167
stage	0.297	0.148	0.045	0.298	0.150	0.048	0.477	0.187	0.011
node	-0.811	0.226	0.000	-0.696	0.301	0.021	-0.747	0.185	0.000

trt1 and trt2 are the two treatment indicators, stage is the clinical stage I indicator, and node is the lymph nodes involvement indicator. The control arm (both trt1 and trt2 equal zero) is treated as baseline in our analysis. BJ: Buckley-James estimator; LBJ: local Buckley-James estimator; WLS: weighted least squares estimator.

the BJ estimates, resulting in smaller p-values across all variables. LBJ and BJ identify all covariates as significant, while WLS fails to identify the significant effectiveness of Treatment 2. Using the same data set, previous analyses by Farewell (1986), Peng and Dear (2000), and Lu (2010) suggested that Treatment 1 is significantly beneficial in short-term survival while Treatment 2 is significantly beneficial in long-term survival, and that clinical stage I is significantly associated with both short-term and long-term survivals. Compared to WLS, the results from the proposed LBJ method are more in line with these analyses.

6. Discussion

In this paper, we developed a new estimation method for the semiparametric accelerated failure time model with heteroscedastic random errors. Compared with most existing methods, our proposed method is more flexible, and it allows both heteroscedasticity and covariate-dependent censoring. The proposed LBJ estimator does not require estimating the nonparametric heteroscedasticity function $\sigma(\cdot)$, which is computationally appealing. To further improve the efficiency of the LBJ estimator, one might consider a weighted least squares estimation approach by incorporating the nonparametric estimation of $\sigma(\cdot)$.

Motivated by survival studies and by the generalized linear model, we assumed that the error variance is related to the linear mean survival function $X_i^T\beta$ through a nonparametric link function $\sigma(\cdot)$. The essence of the proposed idea can be adapted for models with more general forms of heteroscedasticity, for instance, by allowing the conditional variance to depend on the index $X_i^T\gamma$ with γ a p-dimensional vector possibly different from β. Further research is needed in this direction.

Acknowledgement

We thank the Editor, an associate editor, and two reviewers for their constructive comments to improve the paper. We acknowledge support from the National Institutes of Health grant RO1 CA140632 and National Science Foundation Career Award DMS-1149355.

References

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

Chen, S. and Khan, S. (2000). Estimating censored regression models in the presence of nonparametric multiplicative heteroscedasticity. *J. Econometrics* **98**, 283–316.

Dabrowska, D. (1987). Non-parametric regression with censored survival time data. *Scand. J. Statist.* **14**, 181-197.

Fan, J. and Gijbels, I. (1994). Censored regression: Local linear approximations and their applications. *J. Amer. Statist. Assoc.* **89**, 560-570.

Farewell, V. (1986). Mixture models in survival analysis: Are they worth the risk? *Canad. J. Statist.* **14**, 257-262.

Heuchenne, C. and Van Keilegom, I. (2007). Polynomial regression with censored data based on pre-liminary nonparametric estimation. *Ann. Inst. Statist. Math.* **59**, 273-297.

Jin, Z., Lin, D. Y., Wei, L. J. and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90**, 341.

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

Koul, H., Susarla, V. and Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *Ann. Statist.* **9**, 1276-1288.

Lai, T. L. and Ying, Z. (1991). Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. *Ann. Statist.* **19**, 1370-1402.

Leurgans, S. (1987). Linear models, random censoring, and synthetic data. *Biometrika* **74**, 301-309.

Liu, W. and Lu, X. (2009). Weighted least squares method for censored linear models. *J. Nonparametr. Stat.* **21**, 787-799.

Lu, W. (2010). Efficient estimation for an accelerated failure time model with a cure fraction. *Statist. Sinica* **20**, 661-674.

Nadaraya, E. (1964). On estimating regression. *Theory Probab. Appl.* **9**, 141-142.

Peng, Y. and Dear, K. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics* **56**, 237-243.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *Ann. Statist.* **18**, 303-328.

Stare, J., Heinzl, H. and Harrell, F. (2000). On the use of Buckley and James least squares regression for survival data. *New Approaches in Appl. Statist.* **16**, 125-134.

Tsiatis, A. (1990). Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.* **18**, 354-372.

Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *Ann. Statist.* **21**, 76-99.

Zeng, D. and Lin, D. Y. (2007). Efficient estimation for the accelerated failure time model. *J. Amer. Statist. Assoc.* **102**, 1387-1396.

Zhang, M. and Davidian, M. (2008). "Smooth" semiparametric regression analysis for arbitrarily censored time-to-event data. *Biometrics* **64**, 567-576.

Zhou, M. (2005). Empirical likelihood analysis of the rank estimator for the censored accelerated failure time model. *Biometrika* **92**, 492-498.

Zhou, M., Bathke, A. and Kim, M. (2012). Empirical likelihood analysis for the heteroscedastic accelerated failure time model. *Statist. Sinica* **22**, 295-316.

Zhou, M. and Li, G. (2008). Empirical likelihood analysis of the Buckley-James estimator. *J. Multivariate Anal.* **99**, 649-664.

Merck & Co., Inc, North Wales, PA 19454, U.S.A.

E-mail: panglei.1983@gmail.com

Department of Statistics, North Carolina State University, Raleigh, NC 27695, U.S.A.

E-mail: lu@stat.ncsu.edu

Department of Statistics, George Washington University, Washington, D.C. 20052, U.S.A.

E-mail: judywang@gwu.edu

(Received October 2013; accepted June 2014)