Summer 7-15-2011

Jackknife Empirical Likelihood for the Accelerated Failure Time Model with Censored Data

Maxime K. Bouadoumou
Georgia State University

JACKKNIFE EMPIRICAL LIKELIHOOD FOR THE ACCELERATED FAILURE TIME MODEL WITH CENSORED DATA

by

MAXIME BOUADOUMOU

Under the Direction of Dr. Yichuan Zhao

ABSTRACT

Kendall and Gehan estimating functions are used to estimate the regression parameter in accelerated failure time (AFT) model with censored observations. The accelerated failure time model is the preferred survival analysis method because it maintains a consistent association between the covariate and the survival time. The jackknife empirical likelihood method is used because it overcomes computation difficulty by circumventing the construction of the nonlinear constraint. Jackknife empirical likelihood turns the statistic of interest into a sample mean based on jackknife pseudo-values. *U*-statistic approach is used to construct the confidence intervals for the regression parameter. We conduct a simulation study to compare the Wald-type procedure, the empirical likelihood, and the jackknife empirical likelihood in terms of coverage probability and average length of confidence intervals. Jackknife empirical likelihood method has a better performance and overcomes the under-coverage problem of the Wald-type method. A real data is also used to illustrate the proposed methods.

INDEX WORDS: Confidence interval, Coverage probability, Jackknife empirical likelihood, Right-censoring, U-statistic, Kendall's estimating equation Gehan, Logrank

JACKKNIFE EMPIRICAL LIKELIHOOD FOR THE ACCELERATED FAILURE TIME MODEL WITH CENSORED DATA

by

MAXIME BOUADOUMOU

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2011

JACKKNIFE EMPIRICAL LIKELIHOOD FOR THE ACCELERATED FAILURE TIME MODEL WITH CENSORED DATA

by

MAXIME BOUADOUMOU

Committee Chair: Dr.Yichuan Zhao

Committee: Dr. Jun Han

Dr. Yuanhui Xiao

Electronic Version Approval:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
August 2011

ACKNOWLEDGEMENTS

This thesis would not have been possible without the support of all the people who lent me their supports in different ways. I would like to express my deepest gratitude and sincere appreciation to each and every of them. First of all, I would like to thank my advisor Dr. Yichuan Zhao for sharing his great wealth of knowledge in statistics, especially in survival analysis with me. His encouragement, insightful guidance and his patience with me are highly appreciated and always remembered.

Second of all, I would also like to thank my committee members Dr. Jun Han and Dr. Yuanhui Xiao for taking some valuable time out of their schedules to read my thesis and for giving me some critical suggestions in my thesis research.

I would also want to thank all the faculty members in the department of statistics for teaching me and helping me develop my statistical knowledge throughout my graduate study.

Thanks to all my classmates who helped me through these years and made it all bearable. Without the help of everyone, I would by no means be able to complete this research paper.

TABLE OF CONTENTS

List of Tables

Chapter 1

INTRODUCTION

1.1 Accelerated Failure Time Model

The Cox (1972) proportional hazards model is a popular survival analysis method used to establish a relationship between the covariate and the survival time in censored data. But in many cases, the Cox (1972) proportional hazards model does not always lead to a consistent estimate of the variance and the parameter when the assumptions are not satisfied. To maintain the consistency between the survival time and the covariates, an alternative method called accelerated failure time (AFT) model is quite popularly used. The AFT model assumes that the effect of a covariate is to multiply the predicted event time by some constant. AFT models can be therefore framed as linear models for the logarithm of the survival time. In recent years, many statisticians proposed different estimating methods for the accelerated failure time (AFT); among them, Tsiatis (1990), Ying (1993), and Ritov (1990). These researchers developed an estimating equation based on the linear rank test. Although in theory this method is useful in determining some statistics, it encountered some difficulties when the rank estimation equation is not monotone or continuous. Other researchers such as Lin et al. (1998) developed a root-finding technique called the linear programming method used to estimate the parameters. To estimate the variance, Tsiatis (1990) developed the nonparametric density function. Wei et al. (1990) also developed a method to estimate a consistent variance under certain conditions. Parzen et al. (1994) estimate the limiting covariance matrices by using a re-sampling method. Although these methods are useful to estimate the variance, they still fail to overcome the under-coverage problem of traditional method. Monotone estimating functions based on Kendall and Gehan estimating equations are used in this thesis.

1.2 Monotone Gehan estimating equation

Fygenson and Ritov (1994) originally developed the rank-based estimating equation for right censoring data. Statisticians such as Lin et al. (1998), inspired by Fygenson and Ritov (1994) estimating equation, used the linear programming technique to find a consistent root. Although the latest method is theoretically useful in some cases, in practice it generally fails to estimate the variance of estimator because the equation is not differentiable. In order to overcome the problems mentioned above, Zhao (2011) proposed EL method. We will apply the jackknife empirical likelihood (JEL) developed by Jing, Yuan, and Zhou (2009) to the monotone Gehan estimating equation used in Zhao (2011) to get a better interval estimation of regression parameters.

1.3 Kendall estimating equation

The Kendall's rank regression estimate is defined as follows

$$U_n(b) = \{N \binom{n}{2}\}^{-1} \sum_{1 \le i < j \le n} sgn(X_j - X_i) sgn(\varepsilon_j(b) - \varepsilon_i(b)) \qquad (1)$$

where $\varepsilon_j(b) = Y_j - \mathrm{b}X_j$ and N is $\sum_{1 \le i < j \le n} sgn(X_j - X_i)$.

Kendall estimate is robust against the covariate outliers, as is the Gehan estimating equation, but it is not differentiable in β and required a new method to estimate the asymptotic variance of the regression estimate. Lu (2009) developed EL for the AFT model based on Kendall's estimating equation. In this thesis, we will use the jackknife empirical likelihood to construct the confidence regions for the regression parameter based on the Kendall estimating equation.

1.4 Empirical Likelihood

Empirical likelihood (EL) method was first introduced by Owen (1988, 1990) to determine the shape of the confidence regions without having to estimate the variance. The empirical likelihood does not assume a parametric family of distributions for the data. The empirical likelihood method has been extended in different fields such as on the two-sample problems (Liu, Zou and Zhang (2008), Shen and He (2007)), censored median regression model (Zhao and Chen (2008), Zhao and Yang (2008)) etc. For censored linear regression model including AFT models, recent work of EL includes Zhou (2005), Zhao and Huang (2007), and Zhou and Li (2008), Zhao (2011) etc. For a more thorough review of EL before 2001, you may read Owen (2001).

Based on Fygenson and Ritov (1994) estimating equation, Zhao (2011) developed an EL method for the AFT model. Zhao (2011) developed a procedure that avoids the estimation of the variance for normal approximation based method. Motivated by Subramanian (2007), a profile EL for any specified q components of regression parameters is proposed and by using EL, the $4\chi_q^2$ limiting distribution of the proposed profile El ratio is obtained accordingly. For more discussions on the EL ratio for p-dimension regression analysis, please see Zhao (2011).

1.5 Jackknife Empirical Likelihood

Empirical likelihood is very useful in many different occasions, particularly when data subjects to constraints are linear. However, when applied to more complicated statistics such as U-statistics, it runs into serious computational difficulties. To overcome these difficulties, Jing, Yuan and Zhou (2009) proposed the jackknife empirical likelihood (JEL) for a U-statistic. The method combines two of the popular nonparametric approaches: the jackknife and the empirical likelihood. The key idea of the JEL is to turn the statistic of interest into a sample mean based on jackknife pseudo-values (Quenouille, 1956). If we can show that these pseudo-values are asymptotically independent, we can apply Owen's empirical likelihood for the mean of the jackknife pseudo-values. The most attractive feature with the JEL method is its *simplicity*, as it is merely a simple application of Owen's empirical likelihood to the "sample" mean of jackknife pseudo-values. Theoretically, we will establish Wilks' theorem for one and two-sample *U*-statistics alike. This indicates that the JEL might be potentially useful in handling more general class of statistics than *U*-statistics. Finally, the simulation studies indicate that the JEL compares favorably with other alternatives, and is worthy of serious considerations in statistical inference due to its simplicity.

1.6 Brief History

Jackknife empirical likelihood (JEL) is based on both the jackknife and empirical likelihood methods, and can work in rather general settings beyond the simple i.i.d. settings. JEL can also work under weak assumptions so as to make it as widely applicable as possible. JEL works for one and two-sample *U*-statistics. The two samples can be independent but not identically distributed. For other nonlinear statistics, the validity of the JEL has to be checked case by case. The procedure is as follows. For a defined *U*-statistic, we construct a jackknife sample (see, e.g., Shao and Tu (1995)) first, and then treat this jackknife pseudo sample as a sample of i.i.d. observations and apply the standard empirical likelihood method for the mean of i.i.d. observations to obtain the empirical likelihood ratio statistic for the *U*-statistic.

The empirical likelihood method is one of the most famous methodologies for nonparametric statistical inference procedure which has excellent properties. The deployment of empirical likelihood method with respect to survival analysis can be traced back to Thomas and Grunkemeier (1975). The empirical likelihood method was summarized and discussed in Owen (1988, 1990, 1991), by introducing many great applications and extensions such as constructing nonparametric confidence intervals. Subsequently, Owen and many other statisticians developed this method into a general methodology.

Jackknife techniques have a long history in statistics. The jackknife method of bias reduction was originally proposed by Quenouille (1956), and then Tukey (1958) subsequently demonstrated how the method could be used to construct a nonparametric estimator of variance. As result, it is often referred to as the Quenouille-Tukey jackknife; see, for example, Efron (1982, p.1). According to Miller (1964, p.1594) the procedure was named the jackknife by Tukey because “a boy scout's jackknife is symbolic of a rough-and-ready instrument capable of being utilized in all contingencies and emergencies." The idea behind the jackknife method of bias reduction is to combine a statistic based on a full sample of data with a set of statistics based on sub-samples in a way that eliminates the first-order bias term from its expectation. The interest is often an estimator of a parameter or parameter vector although functions of model parameters and test statistics can also be considered provided they satisfy (or are assumed to satisfy) certain properties. In the case of a random sample of (i.i.d.) variables the sub-samples are usually obtained by deleting observation i from the full sample. This is sometimes known as the delete-1 jackknife because each sub-sample deletes one observation at a time.

The rest of the thesis is organized as follows. The jackknife empirical likelihood is proposed in Chapter 2 and we also present the procedure related to the methods. Three simulation scenarios performances are presented in Chapter 3: the jackknife empirical likelihood confidence intervals, the empirical likelihood and the traditional normal approximation confidence intervals. We then compare the JEL for Gehan and the Kendall’s estimating equations to EL based on Buckley-James, logrank and Gehan estimators, by Zhou (2005) and Zhou and Li (2008). In Chapter 4, we present a real data application. In Chapter 5, a discussion is made and all the technical derivations such as the tables of simulation results and the table of the real application results are presented. The Matlab codes are provided in the Appendix.

Chapter 2

INFERENCE PROCEDURE

2.1 Preliminaries

In this sequel, we use the same notations as those in Lu (2009). The setting of the AFT model is as follows. For i = 1,…,n, let T_i be the failure time for *ith* patient and let Z_i's be the associated $(p \times 1)$ vectors of covariates sequence. $T_i > 0$, representing the time to event in survival analysis. The AFT model is to relate the regression of the logarithm of survival times, $logT_i$, to their p covariates through a standard linear regression equations,

$$logT_i = \beta' Z_i + \epsilon_i, i = 1, \ldots, n \tag{2}$$

where the stochastic errors ϵ_i are independent identically distributed with unknown distribution function F and the covariate vector Z_i is independent of ϵ_i . Since F is unknown, an estimating equation is a natural approach for estimation and inference on β . We assume that C_i is the censoring times for T_i . Assuming that C_i and T_i are independent conditionally on Z_i, we can only observe $X_i = min(T_i, C_i), \Delta_i = I(T_i < C_i)$ and $Z_i,\ i = 1, \ldots, n,$ where $I(\cdot)$ is an indicator function. We assume that ϵ_i is independent of (C_i, Z_i) as was in Fygenson and Ritov (1994). For the regularity conditions, see p.735 of Fygenson and Ritov (1994) and Zhao (2011).

Using the AFT model, we apply the Kendall and Fygenson-Ritov estimating equations to build a confidence region. We use an estimator introduced by Fygenson and Ritov as a solution of a monotone estimating equation. Assume that β_0 is the true value of β, we obtain the following equation $\varepsilon_i(\beta) = log\ X_i - \beta o^T Z_i$. The Kendall estimating equation is written as

$$U(\beta) = n^{-3/2} \sum_{i=1}^{n} \sum_{j=1}^{n} sgn(Z_i - Z_j) \Delta_i I \{e_j(\beta) > e_i(\beta)\} \quad (3)$$

where *sgn (·)* is a sign function defined as $sgn(t) = \begin{cases} 1 & ,t > 0 \\ 0 & ,t = 0 \\ -1 & ,t < 0. \end{cases}$

The Kendall estimate is robust against outliers (see Heller, 2007). Based on this method, for any fixed β, $U(\beta)$ becomes a simple U-statistic; we can rewrite it as a U-statistic with symmetric kernel,

$$U(\beta) = n^{-\frac{3}{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} sgn(Z_i - Z_j)[\Delta_i I \{e_j(\beta) > e_i(\beta)\} - \Delta_j I\{e_i(\beta) > e_j(\beta)\}]. \quad (4)$$

Assume the same conditions as above, the Gehan estimating equation is written as follows

$$U(\beta) = n^{-3/2} \sum_{i=1}^{n} \sum_{j=1}^{n} (Z_i - Z_j) \Delta_i I\{e_j(\beta) > e_i(\beta)\}. \quad (5)$$

Based on this method, for any fixed β, $U(\beta)$ is a simple U-statistic; we can rewrite it as a U-statistic with symmetric kernel,

$$U(\beta) = n^{-\frac{3}{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} (Z_i - Z_j)[\Delta_i I \{e_j(\beta) > e_i(\beta)\} - \Delta_j I\{e_i(\beta) > e_j(\beta)\}]. \quad (6)$$

Please refer to Zhao (2011) for more discussions on how to determine the asymptotic variance and the confidence interval of the Wald-type procedure.

As stated in Zhao (2011), although the Wald-type estimation method has its excellent properties, it suffers a serious under-coverage problem for a small sample. Empirical likelihood is therefore used to fix the under-coverage problem for linear constraints but when the applications involve nonlinear statistics, EL loses its computational appeals. Jing, Yuan, and Zhou (2009) proposed a new method called jackknife empirical likelihood to overcome the computational burdens.

2.2 The JEL confidence region/interval

Let us consider the jackknife empirical likelihood approach in order to make the computation more appealing. By using the Kendall estimating equation, we have the following

$U_i = (Z_i, X_i, \Delta_i)$ and $h\left(U_i, U_j; \beta\right) = sgn\left(Z_i - Z_j\right)\left\{\Delta_i I\left(e_j(\beta) > e_i(\beta)\right) - \Delta_j I\left(e_i(\beta) > e_j(\beta)\right)\right\}$. (7)

A *U*-statistic of degree *2* with a symmetric kernel *h* is defined to be

$$U_n(\beta) = \binom{n}{2}^{-1} \sum_{j=1, j \neq i}^{n} \{h(U_i, U_j; \beta)\}. \qquad (8)$$

Applying the JEL of Jing et al. (2009) to the above equations, we obtain $U_n(\beta) = U(U_1, \dots, U_n)$, and the jackknife pseudo-values is defined as

$$\hat{V}_i = nU_n - (n-1)\, U_{n-1}^{(-i)}, \qquad (i = 1, \dots, n)$$

where $U_{n-1}^{(-i)} := U\,(U_1, \dots, U_{i-1}, U_{i+1}, \dots, U_n)$, which is obtained from the original data set by removing the *ith* data value.

Also, similarly, from p.736 of Fygenson and Ritov (1994), $E\ddot{V}_i = 0$, hence the sample mean is as follows

$$\sum_{i=1}^{n} p_i E\hat{V}_i = 0.$$

Furthermore, the jackknife estimator is defined as

$$U_n = \frac{1}{n}\sum_{i=1}^{n} \hat{V}_i.$$

Generally, the jackknife pseudo values are r.v.'s, but asymptotically independent under weak or mild conditions (see Shi, 1984). We can then apply the JEL to the jackknife pseudo values. Let $p = (p_1, \ldots, p_n)$ be a probability vector. Then the empirical likelihood function at the value β is given by

$$L(\beta) = max\left\{\prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \hat{V}_i = 0, p_i \geq 0\right\}.$$

Note that $\prod_{i=1}^{n} p_i$ attains its maximum at $p_i = 1/n$. Thus, the jackknife empirical likelihood ratio at β is defined by $R(\beta) = \frac{L(\beta)}{n^{-n}}$,

$$R(\beta) = max\left\{\prod_{i=1}^{n} np_i : \sum_{i=1}^{n} P_i = 1, \sum_{i=1}^{n} p_i \hat{\boldsymbol{V}}_i = 0, p_i \geq 0\right\} \qquad (9)$$

and its logarithm form is

$$R(\beta) = max\left\{\sum_{i=1}^{n} \log(np_i) : \sum_{i=1}^{n} P_i = 1, \sum_{i=1}^{n} p_i \hat{V}_i = 0, p_i \geq 0\right\}. \quad \mathbf{(10)}$$

By using the Lagrange multipliers method,

we have

$$p_i = \frac{1}{n}\frac{1}{1 + \lambda(\hat{V}_i - E\hat{V}_i)}$$

where λ satisfies

$$f(\lambda) \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{V}_i - E\hat{V}_i}{1 + \lambda(\hat{\boldsymbol{V}}_i - E\hat{V}_i)} = 0.$$

Then, by plugging p_i into the logarithm transformation of $R(\beta)$, we obtain

$$\hat{l}(\beta) = -2logR(\beta) = -2\sum_{i=1}^{n} log(np_i) = -2\sum_{i=1}^{n} log\{1+\lambda(\hat{V}_i)\}. \quad (11)$$

The following theorems establish how Wilks' theorem holds and state how the result can be used to construct confidence region for β.

Theorem 1 *Under the above conditions $\hat{l}(\beta_0)$ converges in distribution to χ_p^2, where χ_p^2 is a chi-square random variable with p degrees of freedom.*

With Theorem 1, an asymptotic 100(1-α) % confidence region for β is given by

$$R_E = \{\beta: \ddot{l}(\beta_0) \leq \chi_p^2(\alpha)\}, \quad (12)$$

where $\chi_p^2(\alpha)$ is the upper α-quantile of the distribution of χ_p^2.

The confidence region for the full set of parameter provides less information in multi-dimensional setting. Regarding the P-values and the confidence intervals for the components of the regression parameters, statisticians make inferences about each element of β. In this thesis what we want to construct is the EL confidence region for the q sub-vector $\beta^{(1)}$ of $\beta = (\beta^{(1)\prime}, \beta^{(2)\prime})'$. Based on Subramanian (2007) profile empirical likelihood for censored median regression models, Zhao (2011) proposed the profile EL for single components by profiling out the nuisance parameters from the full EL. Thus, we will adapt these methods to our settings.

Define $\beta_0 = ({\beta_0}^{(1)\prime}, {\beta_0}^{(2)\prime})'$ and $\mathcal{N} = \{\beta^{(2)\prime}, \| \beta^{(2)\prime} - \hat{\beta}^{(2)\prime} \| = O(n^{-1/3})\}$. The profile EL ratio at $\beta^{(1)}$ is defined as $\hat{l}_{Profile}(\beta^{(1)}) = min_{\beta^{(2)} \in \mathcal{N}} \hat{l}((\beta^{(1)\prime}, \beta^{(2)\prime})')$. The corresponding theorem for the full EL is therefore obtained.

Theorem 2: *Under the above conditions, $\hat{l}_{Profile}(\beta_0^{(1)})$ converges in distribution to χ_q^2, where χ_q^2 is a chi-square random variable with q degrees of freedom.*

Using this Theorem, an asymptotic *100(1-α) %* confidence region for $\beta^{(1)}$ is given by

$$R_P = \left\{\beta_1^{(1)} : \hat{l}_{Profile}(\beta^{(1)}) \leq \chi_q^2(\alpha)\right\}, \qquad (13)$$

where $\chi_q^2(\alpha)$ is the upper α-quintile of the distribution of χ_q^2.

For Fygenson-Ritov estimating equation, similarly we can obtain the profile jackknife EL $\hat{l}_{Profile}(\beta_0^{(1)})$ by eliminating the nuisance parameters from the full JEL. The resulting JEL confidence intervals for $\beta_0^{(1)}$ is obtained, which is consistent with Theorem 1 and Theorem 2.

Chapter 3

SIMULATION STUDY

3.1 JEL, and EL vs. Wald-type based on Gehan and Kendall estimating equations

Based on the Gehan and Kendall estimating equations, extensive simulation studies are conducted to compare the performances of the confidence intervals of jackknife empirical likelihood, empirical likelihood method and the Wald-type based approach. The performances of the proposed procedures are compared in terms of coverage probability and average length of confidence intervals in different settings. We use the same settings as those in Lu (2009).

Assuming there are only one covariate Z and a true parameter $\beta_0 = 2$, skewed error distribution and the symmetric error distribution are the two models considered to conduct the simulation runs. Model 1 has a covariate Z and is uniformly distributed in [-1, 1]. The censoring time C follows uniform distribution in [0, c], where c controls the censoring rate. The error term has a standard Gumbel distribution when $\mu = 0$ and $\beta = 1$, and the cumulative distribution function skewed to the right is defined as follows

$$F(x) = e^{-e^{-x}}.$$

We generated the error term as the following $\varepsilon_i = log(-log(U))$, where U is a uniform variable in [0, 1]. Thus, the survival time can be obtained by $T_i = exp(\beta_0 Z_i + \varepsilon_i)$, where $\beta_0 = 2$.

Model 2 has a covariate Z and is uniformly distributed in [0.5, 1.5].The censoring time C is distributed as 2 exp (1) +c, where exp (1) is a standard exponential distribution, and c controls the censoring rate. The symmetric error distribution is similar to that of the standard Normal distribution N (0, 1).

The censoring time C can be generated as follows $C = -2log(U) + c$, where *U* is a uniform variable in [0, 1] and c is a constant.

Assuming a true value $\beta_0 = 2$, four different censoring rates with approximately 15%, 30% and 45%, and 60% respectively, which represent light censoring, medium censoring and moderate heavy censoring, and very heavy censoring rate. The sample sizes are 30, 50, 75 and 100, representing very small, relatively small, moderate and large samples respectively. Therefore, we have 16 data settings in total for each of the two models. Each data set is simulated 10000 times and the results are displayed in Table 1 and Table 2.

Table 1 and Table 2 displayed the results of the Wald-type, the empirical likelihood, and the jackknife empirical likelihood methods. The censoring rates are approximately 15%, 30%, 45%, and the sample size, based on 10,000 simulated data sets, is 30, 50, 75, and 100. The three methods have better performances in term of coverage probabilities and average lengths when the total sample size increases. The coverage probability for large sample, that is n=100, works well with right coverage probability of 90%, 95%. The Wald-type method has greater under-coverage when the sample size is small, while the empirical likelihood and the jackknife empirical likelihood methods have a better coverage probability for all the nominal levels. The three methods have a better accuracy of the coverage probabilities when the censoring rate decreases because there are fewer information losses.

Table 1: Coverage probability and average lengths of confidence intervals for the regression parameter β with model 1.

CR	n		1-α=0.90				1-α=0.95			
			Wald	**EL1**	**JEL1**	**JEL2**	**Wald**	**EL1**	**JEL1**	**JEL2**
15%	30	Coverage	0.8691	0.8992	0.8997	0.8989	0.9228	0.9436	0.9489	0.9478
		Length	1.4349	1.5739	1.5741	1.5738	1.7101	1.9037	1.9056	1.9051
	50	Coverage	0.8862	0.9087	0.9098	0.9089	0.9342	0.9524	0.9571	0.9562
		Length	1.0859	1.1844	1.1848	1.1851	1.2943	1.4212	1.4227	1.4223
	75	Coverage	0.8889	0.9134	0.9198	0.9189	0.9411	0.9592	0.9598	0.9593
		Length	0.8735	0.9515	0.9532	0.9527	1.0412	1.1422	1.1431	1.1428
	100	Coverage	0.8941	0.9157	0.9176	0.9163	0.9429	0.9617	0.9669	0.9658
		Length	0.7517	0.8073	0.8086	0.8083	0.8965	0.9783	0.9799	0.9793
30%	30	Coverage	0.8676	0.8953	0.8991	0.8979	0.9168	0.9369	0.9476	0.9468
		Length	1.6869	1.8169	1.8183	1.8177	2.0109	2.1744	2.1753	2.1748
	50	Coverage	0.8769	0.8987	0.8998	0.8989	0.9276	0.9457	0.9489	0.9481
		Length	1.2683	1.3631	1.3645	1.3639	1.5114	1.6262	1.6276	1.6261
	75	Coverage	0.8836	0.9067	0.9103	0.9098	0.9375	0.9521	0.9581	0.9573
		Length	1.0223	1.1047	1.1053	1.1049	1.2169	1.3139	1.3153	1.3148
	100	Coverage	0.8919	0.9117	0.9189	0.9179	0.9422	0.9596	0.9617	0.9609
		Length	0.8816	0.9469	0.9478	0.9471	1.0488	1.1363	1.1404	1.1401
45%	30	Coverage	0.8497	0.8729	0.8789	0.8778	0.9087	0.9194	0.9215	0.9208
		Length	2.0332	2.1771	2.1789	2.1785	2.4218	2.5978	2.6012	2.6007
	50	Coverage	0.8708	0.8852	0.8971	0.8963	0.9238	0.9337	0.9399	0.9388
		Length	1.5253	1.5963	1.6617	1.6609	1.8163	1.9011	1.9084	1.9079
	75	Coverage	0.8805	0.9011	0.9089	0.9081	0.9342	0.9441	0.9488	0.9479
		Length	1.2298	1.2965	1.2974	1.2969	1.4651	1.5283	1.5291	1.5286
	100	Coverage	0.8876	0.9046	0.9089	0.9077	0.9399	0.9478	0.9518	0.9512
		Length	1.0521	1.1223	1.0370	1.0369	1.2523	1.3241	1.1897	1.1895
60%	30	Coverage	0.8142	0.8386	0.8678	0.8669	0.8763	0.8868	0.8991	0.8983
		Length	2.6108	2.7791	2.8441	2.8438	3.1101	3.2618	3.3685	3.3681
	50	Coverage	0.8489	0.8494	0.8789	0.8781	0.9013	0.9034	0.9211	0.9203
		Length	1.9466	1.9872	2.1032	2.1028	2.3187	2.3193	2.4893	2.4891
	75	Coverage	0.8712	0.8673	0.8827	0.8821	0.9219	0.9167	0.9376	0.9371
		Length	1.5722	1.5893	1.6645	1.6639	1.8744	1.8752	1.8898	1.8591
	100	Coverage	0.8744	0.8816	0.8975	0.8968	0.9289	0.9279	0.9432	0.9428
		Length	1.3365	1.3826	1.4321	1.4316	1.6032	1.6217	1.7127	1.7118

CR: censoring rate
EL1: empirical likelihood using Kendall estimating equation
JEL1: jackknife empirical likelihood using Kendall estimating equation
JEL2: jackknife empirical likelihood using Gehan estimating equation

Table 2: Coverage probability and average lengths of confidence intervals for the regression parameter β with model 2

			1-α=0.90				1-α=0.95			
CR	n		**Wald**	**EL1**	**JEL1**	**JEL2**	**Wald**	**EL1**	**JEL1**	**JEL2**
15%	30	Coverage	0.8543	0.9086	0.9113	0.9107	0.9126	0.9511	0.9609	0.9598
		Length	2.3071	2.4428	2.5142	2.5133	2.7488	2.8873	2.9251	2.8898
	50	Coverage	0.8758	0.9165	0.9192	0.9183	0.9297	0.9614	0.9689	0.9679
		Length	1.7436	1.8877	1.8886	1.8878	2.0776	2.2515	2.2569	2.2563
	75	Coverage	0.8856	0.9184	0.9196	0.9189	0.9379	0.9632	0.9667	0.9658
		Length	1.4098	1.5139	1.5147	1.5143	1.6798	1.8265	1.8289	1.8283
	100	Coverage	0.8884	0.9124	0.9202	0.9193	0.9419	0.9632	0.9688	0.9682
		Length	1.2162	1.2854	1.3051	1.2984	1.4492	1.5598	1.6012	1.6007
30%	30	Coverage	0.8526	0.9006	0.9117	0.9111	0.9069	0.9462	0.9512	0.9504
		Length	2.4351	2.5434	2.6012	2.6006	2.9013	2.9931	3.1241	3.1232
	50	Coverage	0.8764	0.9064	0.9114	0.9109	0.9239	0.9519	0.9627	0.9621
		Length	1.8432	1.9751	2.1131	2.1127	2.1966	2.3565	2.4126	2.4122
	75	Coverage	0.8831	0.9127	0.9184	0.9179	0.9379	0.9616	0.9669	0.9662
		Length	1.4887	1.5897	1.6117	1.6113	1.7736	1.9086	1.9125	1.9122
	100	Coverage	0.8834	0.9092	0.9124	0.9121	0.9383	0.9588	0.9601	0.9598
		Length	1.2811	1.3529	1.3613	1.3611	1.5259	1.6345	1.7112	1.7109
45%	30	Coverage	0.8441	0.8838	0.8991	0.8983	0.8991	0.9333	0.9464	0.9458
		Length	2.6694	2.8715	2.9254	2.9249	3.1811	3.4432	3.5213	3.5203
	50	Coverage	0.8693	0.8965	0.9013	0.9007	0.9211	0.9419	0.9512	0.9503
		Length	2.0311	2.1351	2.2354	2.2349	2.4198	2.5534	2.6232	2.6228
	75	Coverage	0.8807	0.9081	0.9109	0.9102	0.9331	0.9498	0.9522	0.9515
		Length	1.6303	1.7278	1.7892	1.7885	1.9422	2.0531	2.6211	2.6207
	100	Coverage	0.8878	0.9079	0.9126	0.9121	0.9369	0.9545	0.9598	0.9591
		Length	1.4075	1.4886	1.5125	1.5122	1.6771	1.7754	1.7768	1.7761
60%	30	Coverage	0.8322	0.8636	0.8988	0.8982	0.8871	0.9095	0.9371	0.9362
		Length	3.0165	3.1971	3.4178	3.4171	3.5937	3.7969	3.8112	3.8103
	50	Coverage	0.8708	0.8785	0.8996	0.8991	0.9181	0.9224	0.9463	0.9456
		Length	2.2773	2.3437	2.4214	2.4206	2.7134	2.7911	2.8321	2.8315
	75	Coverage	0.8820	0.8946	0.9087	0.9078	0.9305	0.9398	0.9512	0.9508
		Length	1.8234	1.8976	1.9214	1.9208	2.1725	2.2440	2.4175	2.4171
	100	Coverage	0.8811	0.8996	0.9021	0.9017	0.9332	0.9424	0.9556	0.9549
		Length	1.5664	1.6454	1.7231	1.7227	1.8663	1.9424	2.1231	2.1226

The Wald-type procedure has a slightly shorter average length compared to the empirical likelihood and the jackknife empirical likelihood methods; thus, the shorter the average length of confidence interval, the better the confidence interval. Also we notice that when sample size increases, the average length shortens and the censoring rate decreases and the reason is the larger the sample size, the less information is susceptible to be lost. The Wald type confidence interval is symmetric; however the empirical likelihood and the jackknife empirical likelihood confidence intervals are not symmetric because EL1 and JEL confidence intervals are built through their data set instead of a given distribution. In term of coverage probability, JEL1 is better than JEL2 which is also better than EL1 but in terms of average length, the Wald type method has a slightly shorter length than JEL1 and JEL2.

3.2 Kendall vs. Buckley-James vs. Gehan vs. Logrank Estimator

Jackknife empirical likelihood method is used to compare Kendall's Tau, Buckley-James (Zhou and Li , 2008) and Logrank estimators (Zhou, 2005) in terms of coverage probability and average length. In this regard, model 3 is introduced with a covariate Z and the error term follows a Normal distribution in *N* (1, 0.5^2). We use the same settings as those in Lu (2009).

The censoring time C also follows a Normal distribution in *N*(μ, 4^2), where $\mu = -1.8, 1, 3.1 \text{ } and \text{ } 6.1$ respectively and censoring rates are 10%, 30%, 50%, 75%. The sample size n is 50, 100 and 200. The coverage probabilities are based on 5000 simulation repetitions. Considering $\beta_0 = 1$

we can observe $X_i = \min(Y_i, C_i)$, $\Delta_i = I(Y_i \leq C_i)$

and Z_i, $i = 1, \dots, n$, where $I(.)$ is an indicator function,

the AFT model is defined as

$$Y_i = \beta Z_i + \epsilon_i, \quad i = 1, \dots, n.$$

Table 3: Coverage probability using JEL, EL1, B-J, Logrank, and Gehan estimators

		1-α=0.90						1-α=0.95					
Censoring rate	n	B-J	Logrank	Gehan	EL1	JEL1	JEL2	B-J	Logrank	Gehan	EL1	JEL1	JEL2
10%	50	0.8932	0.8885	0.8844	0.9125	0.9179	0.9163	0.9417	0.9411	0.9367	0.9522	0.9559	0.9546
	100	0.8893	0.8918	0.8916	0.9234	0.9283	0.9274	0.9419	0.9491	0.9452	0.9632	0.9678	0.9664
	200	0.8815	0.9067	0.8951	0.9046	0.9072	0.9083	0.9467	0.9509	0.9451	0.9511	0.9581	0.9572
30%	50	0.8819	0.8878	0.8812	0.9074	0.9083	0.9072	0.9383	0.9368	0.9299	0.9526	0.9581	0.9569
	100	0.8943	0.8893	0.8879	0.9221	0.9271	0.9279	0.9481	0.9422	0.9392	0.9569	0.9592	0.9581
	200	0.8934	0.9147	0.8967	0.9116	0.9162	0.9149	0.9479	0.9631	0.9453	0.9582	0.9594	0.9585
50%	50	0.8843	0.8807	0.8657	0.8981	0.8998	0.8987	0.9332	0.9329	0.9234	0.9372	0.9426	0.9414
	100	0.8934	0.8946	0.8826	0.9082	0.9097	0.9089	0.9425	0.9528	0.9383	0.9532	0.9569	0.9554
	200	0.8957	0.8932	0.8976	0.9146	0.9183	0.9169	0.9487	0.9475	0.9433	0.9611	0.9669	0.9658
75%	50	0.8428	0.8447	0.8041	0.8558	0.8764	0.8753	0.9047	0.9022	0.8635	0.8869	0.9293	0.9282
	100	0.8824	0.8747	0.8548	0.8863	0.8981	0.8972	0.9352	0.9311	0.9125	0.9345	0.9415	0.9403
	200	0.8935	0.8865	0.8776	0.9027	0.9073	0.9058	0.9447	0.9448	0.9366	0.9497	0.9538	0.9528

B-J: EL for Buckley-James estimator by Zhou and Li (2008)
Logrank: EL for logrank estimator by Zhou (2005)
Gehan: EL for Gehan estimator by Zhou (2005)
JEL1: Jackknife empirical likelihood using Kendall estimating equation by Kendall (1938)
JEL2: Jackknife empirical likelihood using Gehan estimating equation by Fygenson and Ritov (1994)

The results displayed in Table 3 show that jackknife empirical likelihood using Kendall estimating equation (JEL1) has a better performance in terms of coverage probability, which is followed closely by jackknife empirical likelihood using Gehan equation (JEL2). In most cases, Kendall's coverage probability (EL1) has a better performance compared to Buckley-James, Gehan, and Logrank estimators; for smaller sample size, Kendall outperforms the three estimators. Gehan has the worst performance coverage among the different methods. In conclusion, JEL1 and JEL2 have similar coverage probability and are better than Logrank, B-J and EL1 estimators.

Chapter 4

APPLICATION

4.1 Introduction

In this section, we apply jackknife empirical likelihood methods to bone marrow transplant procedure described by Klein and Maeschberger (1997) in “survival analysis: techniques for censored and truncated data”. Lu (2009) also used EL for Kendall’s estimating equation to analyze this dataset. We combine these results together.

The preparative regimen used in this study of allogeneic marrow transplants for patients with acute myeloctic leukemia (AML) and acute lymphoblastic leukemia (ALL) was a combination of 16 mg/kg of oral Busulfan (BU) and 120 mg/kg of intravenous cyclophosphamide (Cy). A total of 137 patients (99 AML, 38 ALL) were treated at one of four hospitals: 76 at The Ohio State University Hospitals (OSU) in Columbus; 21 at Hahnemann University (HU) in Philadelphia; 23 at St. Vincent’s Hospital (SVH) in Sydney Australia; and 17 at Alfred Hospital (AH) in Melbourne. The study consists of transplants conducted at these institutions from March 1, 1984, to June 30, 1989. The maximum follow-up was 7 years. There were 42 patients who relapsed and 41 who died while in remission. Twenty-six patients had an episode of acute GVHD, and 17 patients either relapsed or died in remission without their platelets returning to normal levels.

Several potential risk factors were measured at the time of transplantation. For each disease, patients were grouped into risk categories based on their status at the time of transplantation. These categories were as follows: ALL (38 patients), AML low-risk first remission (54 patients), and AML high-risk second remission or untreated first relapse (15 patients) or second or greater relapse or never in remission (30 patients). Other risk factors measured at the time of transplantation included recipient and donor gender (80 and 88 males respectively), recipient and donor cytomegalovirus immune status (CMV) status (68 and 58 positive, respectively), recipient and donor age (ranges 7–52 and 2–56, respectively), waiting time from diagnosis to transplantation (range 0.8–87.2 months, mean 19.7 months), and, for AML patients, their French-American-British (FAB) classification based on standard morphological criteria. AML patients with an FAB classification of M4 or M5 (45/99 patients) were considered to have a possible elevated risk of relapse or treatment-related death.

The censoring rate used in this simulation study is 40.88% and the model is defined as follows

$logT_i = \ \beta Z_i + \epsilon_i$, where $\mathrm{T_i}$ is Time to Death, and $\mathrm{Z_i}$ is the covariate and ϵ_i is the error term .

Four different coefficient estimates β and a covariate are used to construct the confidence intervals. As Table 4 shows, among the covariates, FAB is very significant across all the nominal levels when it is associated to the Wald-type procedure. Among the other three coefficient estimates, covariate called Group is significant at confidence level 0.90. As in Lu (2009), the variable Age is a scaled interaction between patient Age and donor Age, and is defined as Age = (patient Age - 28) $\times$ (donor Age - 28) /100. Age is significant at several confidence levels. Lastly, we notice that the Waiting time is not significant at any confidence levels as shown in Table 4. JEL1 and JEL2 have longer lengths; therefore, their confidence intervals are longer, which is in phase with the results we found in our simulation study.

4.2 Results and Analysis

Table 4: Confidence intervals using JEL1, JEL2, EL1 and Wald-type procedure

		FAB	Group	Age	TimeToTrx
	β	-0.8388	-0.4558	-0.4588	-0.2055
		CI	CI	CI	CI
Wald	0.90	(-1.3483 -0.3292)	(-0.8418 -0.071)	(-0.7791 0.1408)	(-0.4382 0.0271)
	Length	1.0191	0.7708	0.6383	0.4653
	0.95	(-1.4463 -0.2317)	(-0.9157 0.0042)	(-0.8381 -0.0799)	(-0.4826 0.0715)
	Length	1.2146	0.9199	0.7582	0.5541
	0.99	(-1.6372 -0.0409)	(-1.0602 0.1487)	(-0.9574 0.0396)	(-0.5697 0.1585)
	Length	1.5963	1.2089	0.997	0.7282
EL1	0.90	(-1.3728 -0.2543)	(-0.8629 -0.0385)	(-0.9321 -0.1722)	(-0.4908 -0.0136
	Length	1.1185	0.8244	0.7599	0.4772
	0.95	(-1.4936 -0.1244)	(-0.9445 0.0421)	(-1.0505 -0.1089)	(-0.5731 0.0214)
	Length	1.3692	0.9866	0.9416	0.5945
	0.99	(-1.6192 0.1329)	(-1.1252 0.2236)	(-1.2551 0.0057)	(-0.7335 0.0777)
	Length	1.7521	1.3488	1.2608	0.8112
JEL1	0.90	(-1.4972 -0.2942)	(-0.9275 -0.0369)	(-1.0262 -0.1843	(-0.5308 -0.0214)
	Length	1.2030	0.8906	0.8419	0.5094
	0.95	(-1.5238 -0.1348)	(-1.1563 0.0572)	(-1.3081 -0.1068)	(-0.6339 0.0172)
	Length	1.3890	1.2135	1.2013	0.6511
	0.99	(-1.7278 0.1483)	(-1.2457 0.3189)	(-1.3363 0.0198)	(-0.9663 0.0776)
	Length	1.8761	1.5646	1.3561	1.0439
JEL2	0.90	(-1.4915 -0.2833)	(-0.9279 -0.0198)	(-0.9878 -0.1586)	(-0.5425 -0.0198)
	Length	1.2083	0.9081	0.8292	0.5227
	0.95	(-1.5242 -0.1353)	(-1.1558 0.0395)	(-1.0836 -0.1099)	(-0.6428 0.0259)
	Length	1.3889	1.1953	1.1935	0.6687
	0.99	(-1.7365 0.1379)	(-1.2265 0.3142)	(-1.4006 0.0126)	(-0.9789 0.0756)
	Length	1.8744	1.5407	1.4132	1.0545

Chapter 5

CONCLUSION

The recommended coverage probability is the one that that is close enough to the nominal level and the best average length is the shortest one. In terms of probability coverage, JEL1 and JEL2 are close to their corresponding nominal levels. In addition, when the sample sizes are large the coverage probabilities are more accurate and the average lengths are the shortest. The jackknife empirical likelihood using Kendall estimating equation outperformed the EL1 method and the Wald-type procedure. In terms of average length, Wald procedure is the best. We also notice that JEL1 and JEL2 have better coverage probability. When the sample size increases, all the proposed methods have better performances in terms of coverage probabilities and average lengths; However for smaller sample sizes, JEL1 and JEL2 are the best. In addition, it takes less time to compute JEL1 and JEL2. Zhao (2011) proposed EL method to fix the under-coverage problem that is presented in the Wald-type procedure and this method worked well in our simulation study as well. By using JEL1 and JEL2 to simulate the data set, we noticed that for larger sample size, there is over-coverage problem that needs to be addressed. In our future work, we will investigate the use of bootstrap calibration using JEL to fix the over-coverage problem.

REFERNCES

[1] Andrews, D. F. and Herzberg, A. M., Data: A Collection of Problems from Many fields for the student and Research Worker., *Springer*, 1985.

[2]Arvesen, J. N., Jackknifing U-statistics, *Annals of Mathematical Statistics*,Vol. 40, pp. 2076-2100, 1969.

[3] Efron, B., The jackknife, the bootstrap and other resampling plans, *Society of Industrial and Applied Mathematics CBMS-NSF Monographs*, Vol. 38, pp.1, 1982.

[4] Chen, J., Mulayath, A. and Bovas, A., Adjusted empirical likelihood and its properties, *Journal of Computational and Graphical Statistics*, Vol. 17, pp. 426-443, 2008.

[5] Chen, S.X., Empirical likelihood confidence intervals for nonparametric density estimation, *Biometrika*, Vol 83, pp. 329-341, 1986.

[6] Cox, D.R., Regression models and life-tables (with discussion), *J.R. Statist. Soc. B*., Vol. 34, pp.187-220, 1972.

[7] Fygenson,M. and Ritov,Y., Monotone estimating equations for censored data, *Ann. Math. Statistics*, Vol. 22, pp. 732-746, 1994.

[8] Halmos, P. R., The theory of unbiased estimation, *Annals of Mathematical Statistics*, Vol. 17, pp. 34-43, 1946.

[9] Heller, G., Smoothed rank regression with censored data, *J. Amer. Statist. Assoc*., Vol. 102, pp.552-559, 2007.

[10] Hoeding, W., A class of statistics with asymptotically normal distribution, *Annals of Mathematical Statistics*,Vol. 19, pp. 293-325, 1948.

[11] Jing, B. Y., Yuan, J. Q. and Zhou, W., Jackknife empirical likelihood, *Journal of the American Statistical Association*, Vol. 104, pp. 1224-1232, 2009.

[12] Klein, J.P. and Maoeschberger, M.L., Survival analysis techniques for censored and truncated data, *2nd edition, Springer*, 1997.

[13] Koroljuk, V. S. and Borovskich, Yu. V., Theory of U-statistics, *Mathematics and Its Applications*, Vol. 273, Kluwer Academic Publishers, 1994.

[14] Lee, A. J., U-statistics, Theory and Practice, New York: Marcel Dekker,Inc., 1990.

[15] Li, Y. and Wang, Q.H., Empirical likelihood regression analysis for right censored data, *Statist. Sinica*, Vol. 13, pp. 51-68, 2003.

[16] Lin et al., Accelerated failure time models for counting processes, *Biometrika*, Vol. 85, pp. 605-618, 1998.

[17] Lin,D.Y., and Geyer, C.J., Computational methods for semi-parametric linear regression with censored data, *J.Comp.Graph.Statist.*,Vol 1, pp.77-90, 1992.

[18] Liu,Y., Zou,C., and Zhang, R., Empirical likelihood for the two-sample mean problem, *Statistics and Probability Letters*, Vol. 78, pp. 548-556, 2008.

[19] Lu, Y., Empirical likelihood inference for the accelerated failure time model via Kendall estimating equation, *MS. Thesis*, Georgia State University, 2009.

[20] Miller**,** R. G., A trustworthy jackknife**,** Ann. *Math. Stat.*, Vol. 35, pp. 1594-1605, 1964

[21] Owen, A. B., Empirical likelihood for linear models. *The Annals of Statistics,* Vol. 19, pp. 1725-1747, 1991.

[22] Owen, A. B., Empirical likelihood ratio confidence intervals for a single functional, *Biometrika*, Vol. 75, pp. 237-249, 1988.

[23] Owen, A. B., Empirical likelihood ratio confidence regions. *The Annals of Statistics*, Vol. 18, pp. 90-120, 1990.

[24] Parzen, M. I., Wei, L.J. and Ying, Z., A resampling method based on pivotal estimating functions, *Biometrika*, Vol. 81, pp. 341-350, 1994.

[25] Qin.,G and Jing,B.Y, Empirical likelihood for censored linear regression, *Scand. J. Statist*., Vol. 28, pp. 661-673, 2001.

[26] Qin, J. and Lawless, J., Empirical likelihood and general estimating equations, *The Annals of Statistics*, Vol. 22, pp. 300-325, 1994.

[27] Quenouille, M.H. , Note on bias in estimation, *Biometrika*, Vol. 43, pp. 353-360, 1956.

[28] Ren, J., Weighted empirical likelihood in some two-sample semiparametric models with various types of censored data, *Annals of Statistics*, Vol. 36, pp. 147-166, 2008.

[29] Shao, J. and Tu, D., The Jackknife and Bootstrap, *Springer-Verlag*, 1995.

[30] Shi, X, The approximate independence of jackknife pseudo-values and bootstrap methods, *Journal of Wuhan Institute Hydra-Electric Engineering*, Vol.2, pp. 83-90, 1984.

[31] Subramanian, S., Censored median regression and profile empirical likelihood, *Statistical Methodology*, Vol. 4, pp. 493-503, 2007.

[32] Thomas, D. R. and Grunkemeier, G. L.,Confidence interval estimation of survival probabilities for censored data, *J. Amer. Statist. Assoc*. Vol. 70, pp. 865-871, 1975.

[33] Tsiatis,A.A., Estimating regression parameters using linear rank tests for censored data, *Ann. Statist*, Vol. 18, pp. 354-372, 1990.

[34] Tukey, J.W , Bias and confidence in not-quite large samples , *Annals of Mathematical Statistics*, Vol. 29, pp. 614, 1958.

[35] Wei et al., Linear regression analysis of censored survival data based on rank tests, *Biometrika*, Vol. 77, pp. 845-851, 1990.

[36] Wood, A.T.A., Do, K.A. and Broom, N.M., Sequential linearization of empirical likelihood constraints with application to U-statistics, *J. Comput. Graph. Stat*.,Vol. 5, pp. 365-385, 1996.

[37] Ying, Z., A large sample study of rank estimation for censored regression data, *Ann. Statist*., Vol. 21, pp. 76-99, 1993.

[38] Zhao, Y., Empirical likelihood inference for the accelerated failure time model, *Statistics and Probability Letters*, Vol. 81, issue 5, pp. 603-610, 2011.

[39] Zhao, Y., and Chen, F., Empirical likelihood inference for median regression model via nonparametric kernel estimation, *Journal of Multivariate Analysis*, Vol.99, pp.215-231, 2008.

[40] Zhao,Y., and Huang, Y., Test-based interval estimation under the accelerated failure time model, *Communications in Statistics—Simulation and Computation*, Vol.36, pp.593-605, 2007.

[41] Zhao,Y., and Yang S., Testing treatment effect via empirical likelihood using weighted log rank tests, *Statistics and Probability Letters*, Vol.77, pp.1385-1393, 2007.

[42] Zhou.M., Empirical likelihood analysis of the rank estimator for the censored accelerated failure time model, *Biometrika*, Vol. 92, pp.492-498, 2008.

[43] Zhou,M. and Li,G., Empirical likelihood analysis of the Buckley-James estimator, *Journal of Multivariate Analysis*, Vol. 99, pp. 649-664, 2008.

APPENDIX: MATLAB CODES

Matlab code:

```
%%%Model 1: extreme value distribution for error term. Covariate is
%uniformly distributed in [-1,1] and the censoring time is uniform [0,c]
% c controls the censoring rate.
% k is # of obs. in one data set; C is censoring rate; b is the true value
% of the coefficient of covariate.
-------------------------------function for model 1----------------------

function Data=data1(k,C,b)
censor=0;
for i=1:k
    z(i)=(rand(1)-0.5)*2;

c(i)=(rand(1)*C);
Eps(i)=log(-log(rand));
T(i)=exp(z(i)*b+Eps(i));
if T(i) > c(i)
delta(i)=0;
censor=censor+1;
else delta(i)=1;
end
end
z;
T;
%prop=censor/k
for i=1:k
Data(i,1)=log(min(T(i),c(i)));
Data(i,2)=delta(i);
Data(i,3)=z(i);
end
prop=censor/k;


%%%Model 2: Standard Normal distribution for error term. Covariate is
%uniformly distributed in [0.5,1.5] and the censoring time is uniform [0,c]
% c controls the censoring rate.
% k is # of obs. in one data set; C is censoring rate; b is the true value
% of the coefficient of covariate.

 --------------------------------function for model 2----------------------

function Data=data2(k,C,b)
censor=0;
for i=1:k
```

```
    z(i)=(rand(1)+0.5)*1;

c(i)=(-2*log(rand)+C);
W=randn(1, 1);
Eps(i)=W;
T(i)=exp(z(i)*b+Eps(i));
if T(i) > c(i)
delta(i)=0;
censor=censor+1;
else delta(i)=1;
end
end
z;
T;
%prop=censor/k
for i=1:k
Data(i,1)=log(min(T(i),c(i)));
Data(i,2)=delta(i);
Data(i,3)=z(i);
end
prop=censor/k;


------------------------------generate data for model 1----------------------

%generate 2000 data sets
% for Model 1
% Choose C=10 for the censoring rate of 15% 11.25
% Choose C=4 for the censoring rate of 30% 4.2
% Choose C=2 for the censoring rate of 45% 2
% Choose C=0.95 for the censoring rate of 60%
% Choose C=0.42 for the censoring rate of 75%
% four kinds of sample size: 100 75 50 30

for k=1:10000
Data11=data1(100,0.42,2);
for i=1:100;
for j=1:3
Dataa(i,j,k)=Data11(i,j);
end
end
end

save  nAFT115  Dataa
clear;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

for k=1:10000
Data11=data1(100,0.95,2);
for i=1:100
for j=1:3
Dataa(i,j,k)=Data11(i,j);
End
```

```
end
end


save  nAFT130  Dataa
clear;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

for k=1:10000
Data1l=data1(100,2,2);
for i=1:100
for j=1:3
Dataa(i,j,k)=Data1l(i,j);
end
end
end

save  nAFT145  Dataa
clear;

------------------------------generate data for model 2----------------------

%generate 2000 data sets
% for Model 2
% Choose C=22.5 for the censoring rate of 15% 23.15
% Choose C=12 for the censoring rate of 30% 12.5
% Choose C=6.6 for the censoring rate of 45% 6.85
% Choose C=3.8 for the censoring rate of 60%
% Choose C=1.6 for the censoring rate of 75%
% four kinds of sample size: 100 75 50 30

for k=1:10000
Data1l=data2(100,1.6,2);
for i=1:100
for j=1:3
Dataa(i,j,k)=Data1l(i,j);
end
end
end

save  nAFT215  Dataa
clear;

%%%%%%%%%

for k=1:10000
Data1l=data2(100,3.8,2);
for i=1:100
for j=1:3
Dataa(i,j,k)=Data1l(i,j);
end
end
end
```

```
save   nAFT230   Dataa
clear;
%%%%%%%%%
for k=1:10000
Data11=data2(100,6.6,2);
for i=1:100
for j=1:3
Dataa(i,j,k)=Data11(i,j);
end
end
end

save   nAFT245   Dataa
clear;
```

 -function for coverage probability of JEL using Gehan or Kendall estimating method---

```
function  cov=empcovjack(n,fname,GKtyp)
%CCC=5;
clear
cov1=0;
cov2=0;
cov3=0;
K=size(Dataa,3);
load(fname);
count=0;
b=2
%n=100
%for kk=1:K
%Data=data(n,CCC,b);
for kk=1:K
for j=1:n
for k=1:3
Data(j,k)=Dataa(j,k,kk);
end
end

lambda=Lambdajack(b, n, 0.0001,Data,GKtyp);
    ll=1;
    W=0;
    for it=1:n
        for j1=1:n
            switch(GKtyp)
                case('G')
                    W= W+(Data(it,3)-Data(j1,3))*(Data(it,2)* ((Data(j1,1)-
b*Data(j1,3)) > (Data(it,1)-b*Data(it,3)))- Data(j1,2)*((Data(it,1)-
b*Data(it,3)) > (Data(j1,1)-b*Data(j1,3))))/(n-1)/2;
                case('K')
                    W= W+sign(Data(it,3)-Data(j1,3))*(Data(it,2)*
((Data(j1,1)-b*Data(j1,3)) > (Data(it,1)-b*Data(it,3)))-
Data(j1,2)*((Data(it,1)-b*Data(it,3)) > (Data(j1,1)-b*Data(j1,3))))/(n-1)/2;
            End
```

```
        end
    end




for itt=1:n

        Wn=0;
        for j1=1:n
            switch(GKtyp)
                case('G')
                    Wn= Wn+(Data(itt,3)-Data(j1,3))*(Data(itt,2)*
((Data(j1,1)-b*Data(j1,3)) > (Data(itt,1)-b*Data(itt,3)))-


Data(j1,2)*((Data(itt,1)-b*Data(itt,3)) > (Data(j1,1)-b*Data(j1,3))))/(n-
1)/2;
                case('K')
                    Wn= Wn+sign(Data(itt,3)-Data(j1,3))*(Data(itt,2)*
((Data(j1,1)-b*Data(j1,3)) > (Data(itt,1)-b*Data(itt,3)))-
Data(j1,2)*((Data(itt,1)-b*Data(itt,3)) > (Data(j1,1)-b*Data(j1,3))))/(n-
1)/2;
            end
        end

        U=W-Wn;
        WW=W*n-(n-1)*(W-Wn);
        ll=ll*abs(1+lambda*WW);
    end

    L=2*log(ll);

    if L>0
    count=count+1
    end
cov1=cov1+(L <= 2.7055);
cov2=cov2+ (L <= 3.841459);
cov3=cov3+ (L <= 6.6349);
kk
%cov(1)=cov1/count
%cov(2)=cov2/count
%cov(3)=cov3/count
cov=[cov1/kk cov2/kk cov3/kk]

end

count;
cov(1)=cov1/K
cov(2)=cov2/K
cov(3)=cov3/K
```

-Bisection method for average length of JEL using Gehan or Kendall estimating method -

```
function average=empLengthjack(upper,lower,initial,CL,Data,n,GKtyp)

%%%%%%%%%%%%% lower bound for 95% confidence region%%%

Bl=lower;
Bu=initial;
if CL==90;
    Z=1.645^2;
elseif CL==95;
    Z=1.96^2;
elseif CL==99;
    Z=2.576^2;
end

diff=2;

while (diff>0.00001)
    x=(Bl+Bu)/2;
    B=Lambdajack(x, n, 0.0001, Data, GKtyp);
    ll=ones(1,1);
    for it=1:n
        W=0;
        for j1=1:n
            switch(GKtyp)
                case('G')
                    W= W+(Data(it,3)-Data(j1,3))*(Data(it,2)* ((Data(j1,1)-
B*Data(j1,3)) > (Data(it,1)-B*Data(it,3)))- Data(j1,2)*((Data(it,1)-
B*Data(it,3)) > (Data(j1,1)-B*Data(j1,3))))/(n-1)/2;
                case('K')
                    W= W+sign(Data(it,3)-Data(j1,3))*(Data(it,2)*
((Data(j1,1)-B*Data(j1,3)) > (Data(it,1)-B*Data(it,3)))-
Data(j1,2)*((Data(it,1)-B*Data(it,3)) > (Data(j1,1)-B*Data(j1,3))))/(n-1)/2;
            end
        end
    end

    for itt=1:n
        Wn=0;
        for j1=1:n
            switch(GKtyp)
                case('G')
                    Wn= Wn+(Data(itt,3)-Data(j1,3))*(Data(itt,2)*
((Data(j1,1)-B*Data(j1,3)) > (Data(itt,1)-B*Data(itt,3)))-
Data(j1,2)*((Data(itt,1)-B*Data(itt,3)) > (Data(j1,1)-B*Data(j1,3))))/(n-
1)/2;
                case('K')
                    Wn= Wn+sign(Data(itt,3)-Data(j1,3))*(Data(itt,2)*
((Data(j1,1)-B*Data(j1,3)) > (Data(itt,1)-B*Data(itt,3)))-
Data(j1,2)*((Data(itt,1)-B*Data(itt,3)) > (Data(j1,1)-B*Data(j1,3))))/(n-
1)/2;
            end
        end
        U=W-Wn;
```

```
        WW=W*n-(n-1)*(W-Wn);
        ll=ll*abs(1+B*WW);
    end

    L=2*log(ll);


        if  L < Z
            Bu=x;
        else
            Bl=x;
        end
        diff=Bu-Bl;

    end

    bhatl=Bu;


    %%%%%%%%%%%%%%%%%%%%%%upper bound for 95% confidence region%
    Bu=upper;
    Bl=initial;

    diff=2;

    while (diff>0.00001)
        x=(Bl+Bu)/2;
        B=Lambdajack(x, n, 0.0001, Data, GKtyp);
        ll=ones(1,1);
        for it=1:n
            U=zeros(1,1);
            for it=1:n
                W=0;
                for j1=1:n
                    switch(GKtyp)
                        case('G')
                            W= W+(Data(it,3)-Data(j1,3))*(Data(it,2)*
((Data(j1,1)-B*Data(j1,3)) > (Data(it,1)-B*Data(it,3)))-
Data(j1,2)*((Data(it,1)-B*Data(it,3)) > (Data(j1,1)-B*Data(j1,3)))))/(n-1)/2;
                        case('K')
                            W= W+sign(Data(it,3)-Data(j1,3))*(Data(it,2)*
((Data(j1,1)-B*Data(j1,3)) > (Data(it,1)-B*Data(it,3)))-
Data(j1,2)*((Data(it,1)-B*Data(it,3)) > (Data(j1,1)-B*Data(j1,3)))))/(n-1)/2;
                    end
                end

            end
        end

        for itt=1:n
            Wn=0;
            for j1=1:n
                switch(GKtyp)
                    case('G')
```

```
                        Wn= Wn+(Data(itt,3)-Data(j1,3))*(Data(itt,2)*
((Data(j1,1)-B*Data(j1,3)) > (Data(itt,1)-B*Data(itt,3)))-
Data(j1,2)*((Data(itt,1)-B*Data(itt,3)) > (Data(j1,1)-B*Data(j1,3)))))/(n-
1)/2;
                    case('K')
                        Wn= Wn+sign(Data(itt,3)-Data(j1,3))*(Data(itt,2)*
((Data(j1,1)-B*Data(j1,3)) > (Data(itt,1)-B*Data(itt,3)))-
Data(j1,2)*((Data(itt,1)-B*Data(itt,3)) > (Data(j1,1)-B*Data(j1,3)))))/(n-
1)/2;
                end
            end

            U=W-Wn;
            WW=W*n-(n-1)*(W-Wn);
            ll=ll*abs(1+B*WW);
        end

        L=2*log(ll);
        if  L < Z
            Bl=x;
        else
            Bu=x;
        end
        diff=Bu-Bl;

    end

    bhatu=Bu;

    average=bhatu-bhatl;
```

-- function for Jacobtrans of JEL using Kendall or Gehan estimating method --

```
function  x=Jacobtransjack(b,n, Bold,Data,GKtyp)
% load (fname);
%a0=0;
%a1=-0.25;
%n=30;

   x=0;

   W=0;
   for it=1:n

        for j1=1:n
            switch(GKtyp)
                case('G')
                    W= W+(Data(it,3)-Data(j1,3))*(Data(it,2)* ((Data(j1,1)-
b*Data(j1,3)) > (Data(it,1)-b*Data(it,3)))- Data(j1,2)*((Data(it,1)-
b*Data(it,3)) > (Data(j1,1)-b*Data(j1,3)))))/(n-1)/2;
                case('K')
```

```
                    W= W+sign(Data(it,3)-Data(j1,3))*(Data(it,2)*
((Data(j1,1)-b*Data(j1,3)) > (Data(it,1)-b*Data(it,3)))-
Data(j1,2)*((Data(it,1)-b*Data(it,3)) > (Data(j1,1)-b*Data(j1,3)))))/(n-1)/2;
            end
end

end

for itt=1:n
  Wn=0;
        for j1=1:n
            switch(GKtyp)
                case('G')
                    Wn= Wn+(Data(it,3)-Data(j1,3))*(Data(itt,2)*
((Data(j1,1)-b*Data(j1,3)) > (Data(itt,1)-b*Data(itt,3)))-
Data(j1,2)*((Data(itt,1)-b*Data(itt,3)) > (Data(j1,1)-b*Data(j1,3)))))/(n-
1)/2;
                case('K')
                    Wn= Wn+sign(Data(itt,3)-Data(j1,3))*(Data(itt,2)*
((Data(j1,1)-b*Data(j1,3)) > (Data(itt,1)-b*Data(itt,3)))-
Data(j1,2)*((Data(itt,1)-b*Data(itt,3)) > (Data(j1,1)-b*Data(j1,3)))))/(n-
1)/2;
            end
end


U=W-Wn;
WW=W*n-(n-1)*(W-Wn);
x=x- (WW/(1+Bold*WW))^2;

end



%for it=1:n
    %W=0;
 %       for j1=1:n
 % W(it)= W(it)+(Data(it,3)-Data(j1,3))*(Data(it,2)* ((Data(j1,1)-
b*Data(j1,3)) > (Data(it,1)-b*Data(it,3)))- Data(j1,2)*((Data(it,1)-
b*Data(it,3)) > (Data(j1,1)-b*Data(j1,3)))))/(n-1)/2;
 %end
%x=x- (W(it)/(1+Bold*W(it)))^2;
%end


-- function for Score1trans of JEL using Kendall or Gehan estimating method--

function  S=Score1transjack(b,Bold, n, Data, GKtyp)
S=0;
%W=zeros(1,n);

 W=0;
for it=1:n
        for j1=1:n
```

```
            switch(GKtyp)
                case('G')
                    W= W+(Data(it,3)-Data(j1,3))*(Data(it,2)* ((Data(j1,1)-
b*Data(j1,3)) > (Data(it,1)-b*Data(it,3)))- Data(j1,2)*((Data(it,1)-
b*Data(it,3)) > (Data(j1,1)-b*Data(j1,3)))))/(n-1)/2;
                case('K')

                    W= W+sign(Data(it,3)-Data(j1,3))*(Data(it,2)*
((Data(j1,1)-b*Data(j1,3)) > (Data(it,1)-b*Data(it,3)))-
Data(j1,2)*((Data(it,1)-b*Data(it,3)) > (Data(j1,1)-b*Data(j1,3)))))/(n-1)/2;
end
end
end
for itt=1:n
  Wn=0;
        for j1=1:n
            switch(GKtyp)
                case('G')
                    Wn= Wn+(Data(itt,3)-Data(j1,3))*(Data(itt,2)*
((Data(j1,1)-b*Data(j1,3)) > (Data(itt,1)-b*Data(itt,3)))-


Data(j1,2)*((Data(itt,1)-b*Data(itt,3)) > (Data(j1,1)-b*Data(j1,3)))))/(n-
1)/2;
                case('K')
                    Wn= Wn+sign(Data(itt,3)-Data(j1,3))*(Data(itt,2)*
((Data(j1,1)-b*Data(j1,3)) > (Data(itt,1)-b*Data(itt,3)))-
Data(j1,2)*((Data(itt,1)-b*Data(itt,3)) > (Data(j1,1)-b*Data(j1,3)))))/(n-
1)/2;
            end


end
U=W-Wn;
WW=W*n-(n-1)*(W-Wn);
S=S+ (WW/(1+Bold*WW));
end
%for it=1:n
    %W=0;
%         for j1=1:n
 % W(it)= W(it)+(Data(it,3)-Data(j1,3))*(Data(it,2)* ((Data(j1,1)-
b*Data(j1,3)) > (Data(it,1)-b*Data(it,3)))- Data(j1,2)*((Data(it,1)-
b*Data(it,3)) > (Data(j1,1)-b*Data(j1,3)))))/(n-1)/2;
 %end
%S=S+ (W(it)/(1+Bold*W(it)));
%end
```

- function for solving Lambda in JEL using Gehan or Kendall estimating method

```
function  r=Lambdajack(b,n,eps,Data,GKtyp)
%tau=0.25;
%n=30;
%a0=0;
```

```
%a1=-0.25;
%tau=0.25;
%CCC=5;
Bold=0;
while  (norm(Score1transjack(b,Bold,n,Data,GKtyp))>eps)
    Bold=Bold-inv(Jacobtransjack(b, n,Bold,Data,GKtyp)) * ...
        Score1transjack(b,Bold,n,Data,GKtyp);
count=0;

end

r=Bold;
kkkk=norm(Score1transjack(b,Bold,n,Data,GKtyp));
```