Computational Statistics Project

Comparison of three Semi-Parametric Right-Censored AFT Models using Simulation Methods

Andrew Huang

April 2020

Contents

1 Introduction

Time-to-event (TTE) data is one of the major fields of study in survival analysis. For example, in medical studies, clinical researchers are interested in the time to relapse or death of the patient after the surgery. Usually, longer survival time or time to relapse indicates better efficacy in treatment. Nevertheless, not all the survival time can be accurately observed and recorded because patients withdraw or do not develop the event until the end of the study. The Cox proportional hazards (PH) model (Cox, 1972) and the regression model such as accelerated failure time (AFT) model are the two most popular models applied to data with censored observations.

The linear model is one of the most commonly used statistical models. And, the least-squares method is a regular approach to estimate the parameter in the linear model. However, in analyzing survival data, the linear model and the least-squares method is not often used because the exact survival time of the censored observations is unknown. Therefore, the least-squares method cannot be used directly to calculate the survival/failure time for censored observations (Huang & Jin, 2007). A lot of studies has been conducted on the accommodation to the linear regression with censored data; for example, the Buckley-James estimator by Buckley and James (1979), the rank-based estimator by Tsiatis (1990), the synthetic data method by Leurgans (1982), and the M-Estimator by Zhou (1992). This project used simulation methods to generate error terms following various distributions, to examine and compare three models: the improved Backley-James method by Jin et al. (2006), the synthetic data method by Leurgans (1987), and the inverse probability weighted method by Zhou (1992).

1.1 Accelerated Failure Time (AFT) Models

Let T_i and C_i denote the failure time and censoring time for the ith observation, respectively. And, let X_i denote a $p \times 1$ vector of the covariates for the ith observation. We assume that T_i and C_i are independent conditional on X_i. Write $Y_i = \log(T_i)$ and consider a general form of linear regression model

$$Y_i = X_i^T \beta_0 + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where β_0 is the $p \times 1$ parameter vector.

Usually, parameters of a AFT model can be estimated through the maximum likelihood if a known distribution has been specified for the random error. On the other hand, the AFT model is semiparametric if no specific assumption is made for the error distribution (Pang, 2012). The ϵ_is of the semiparametric model are independent and identically distributed with a common but unknown distribution function F. When censored data is taken into account, the failure time of some Y_is would be unknown. We use (Z_i, δ_i, X_i) to denote the data, where $Z_i = min(T_i, C_i)$, $\delta_i = 1_{\{T_i \leq C_i\}}$, and $1_{\{.\}}$ is the indicator function.

2 The Buckley-James Estimator

In the absence of censored data, the regular least-squares method can be used to estimate the parameter by minimizing the objective function

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \alpha - X_i^T\beta)^2 \tag{2}$$

with respect to α and β, where α refers to the mean of the error distribution. The corresponding 'score' function for β_0 is

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - X_i^T\beta) = 0 \tag{3}$$

where $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$. It would not be hard to prove that the estimated parameter has a simple closed-form expression, and the corresponding covariance matrix can be estimated (Jin at el., 2006).

In the presence of censored data, we do not know the exact failure time of the censored observation, e.g. when $\delta_i = 0$. In this case, we cannot use equation (3) to estimate the value of β_0 directly. Buckley & James (1979) proposed a modified version of equation (3) to accommodate the censored data.The idea is to replace each Y_i with $E(Y_i|\tilde{T}_i, \delta_i, X_i)$, which can be approximated by

$$\hat{Y}_i = \delta_i\tilde{Y}_i + (1-\delta_i)\Big[\frac{\int_{e_i(\beta)}^{\infty} u d\hat{F}_\beta(u)}{1-\hat{F}_\beta\{e_i(\beta)\}} + X_i^T\beta\Big]$$

where $\tilde{Y}_i = log(\tilde{T}_i)$, $e_i(\beta) = \tilde{Y}_i - X_i^T\beta$ and $\hat{F}_\beta$ is the Kaplan-Meier estimator of F based on the transformed data $\{e_i(\beta), \delta_i\}(i = 1, ..., n)$, that is

$$\hat{F}_\beta = 1 - \prod_{i:e_i(\beta)<t}\Big(1 - \frac{\delta_i}{\sum_{j=1}^{n} 1_{e_j(\beta)\geq e_i(\beta)}}\Big) \tag{4}$$

Define

$$U(\beta, b) = \sum_{i=1}^{n}(X_i - \bar{X})\{\hat{Y}_i(b) - X_i^T\beta\},$$

or

$$U(\beta, b) = \sum_{i=1}^{n}(X_i - \bar{X})\{\hat{Y}_i(b) - \bar{Y}(b) - (X_i - \bar{X})^T\beta\},$$

where $\bar{Y}(b) = n^{-1}\sum_{i=1}^{n}\hat{Y}_i(b)$. The Buckley-James estimator $\hat{\beta}_{BJ}$ can be calculated by solving the the equation $U(\beta, \beta) = 0$.

2.1 The Improved Buckley-James Estimator

Jin at al. (2006) proposed an approach to improve the estimation of the Buckley-James estimator. They first fix an initial value b to 'linearise' the estimating function, then solve the equation $U(\beta, b) = 0$ for β. Denote the solution of the previous step as $\beta = L(b)$, where

$$L(b) = \Big\{\sum_{i=1}^{n}(X_i - \bar{X})^{\otimes 2}\Big\}^{-1}\Big[\sum_{i=1}^{n}(X_i - \bar{X})\{\hat{Y}_i(b) - \bar{Y}(b)\}\Big]$$

where the notation $\bigotimes$ follows the rule that $a^{\bigotimes 0} = 1$, $a^{\bigotimes 1} = a$ and $a^{\bigotimes 2} = aa^T$. This process leads to the iterative equation,

$$\hat{\beta}_{(m)} = L(\hat{\beta}_{(m-1)}) \quad (m \geq 1). \tag{5}$$

Lai & Ying (1991) proved that $L(b)$ is asymptotically linear in b. For any fixed m, $\hat{\beta}_{(m)}$ would be consistent if a consistent estimator of $\beta_{(0)}$ is selected to be the initial value in equation (5). Also, if $\beta_{(0)}$ is asymptotically normal, $\hat{\beta}_{(m)}$ would also be asymptotically normal. Jin et al. (2003) introduced a consistent and asymptotically normal initial estimator of $\beta_{(0)}$ which is obtained by the rank-based method. The $\hat{\beta}_{(0)}$ is set to the Gehan-type rank estimator $\hat{\beta}_G$, which can be obtained by minimising the convex function

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\{e_i(\beta) - e_j(\beta)\}^-,$$

where $a^- = 1_{a<0}|a|$.

Jin et al.(2006) introduced a resampling procedure to estimate the covariance matrix of $\beta_{(0)}$. The procedure starts with generating n (the number of observations) identical and independent distributed positive random variables $Z_i (i = 1, 2, ..., n)$, satisfying the condition that $E(Z_i) = Var(Z_i) = 1$. In this study, standard exponential distribution is used to generate Z_is. Similar to equation (4), define

$$\hat{F}_b^*(t) = 1 - \prod_{i:e_i(\beta)<t}\Big(1 - \frac{Z_i\delta_i}{\sum_{j=1}^{n} Z_j 1_{e_j(b)\geq e_i(b)}}\Big) \tag{6}$$

and

$$\hat{Y}_i^*(b) = \delta_i\tilde{Y}_i + (1-\delta_i)\Big[\frac{\int_{e_i(b)}^{\infty} u d\hat{F}_b^*(u)}{1 - \hat{F}_b^*\{e_i(b)\}} + X_i^T b\Big] \tag{7}$$

$$L^*(b) = \Big\{\sum_{i=1}^{n} Z_i(X_i - \bar{X})^{\bigotimes 2}\Big\}^{-1}\Big[\sum_{i=1}^{n} Z_i(X_i - \bar{X})\{\hat{Y}_i^*(b) - \bar{Y}^*(b)\}\Big] \tag{8}$$

Similar to the point estimation of $\hat{\beta}_{(m)}$, equation (8) will lead to the iterative equation $\hat{\beta}^*_{(m)} = L^*(\hat{\beta}^*_{(m-1)})$, $m \geq 1$. we set the initial value $\hat{\beta}^*_{(0)}$ for the iterative function be $\hat{\beta}^*_G$, which can be obtained by minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{n} Z_i Z_j \delta_i |e_i(\beta) - e_j(\beta)| + \Big|M - \beta^T \sum_{i=1}^{n}\sum_{j=1}^{n} Z_k Z_l \delta_k (X_l - X_k)\Big|$$

where M is the prespecified extremely large number. By generating random samples of Z_is repeatedly N times and calculating $\hat{\beta}^*_{(k)} (1 \leq k \leq m)$ using the iterative equation on a given sample of Z_is, we can obtain N realizations of $\hat{\beta}^*_{(m)}$, denoted by $\hat{\beta}^*_{(m),j} (j = 1, ..., N)$. Therefore, for each $m \geq 1$, the covariance matrix of $\hat{\beta}_{(m)}$ can be estimated by

$$s^2 = \frac{1}{N-1}\sum_{j=1}^{N}(\hat{\beta}^*_{(m),j} - \bar{\beta}^*_{(m)})(\hat{\beta}^*_{(m),j} - \bar{\beta}^*_{(m)})^T \tag{9}$$

where $\bar{\beta}^*_{(m)} = (1/N)\sum_{j=1}^{N}\hat{\beta}^*_{(m),j}$.

3 Synthetic Data

Assuming the censoring time C_i does not depend on covariate X_i, Koul et al. (1981) developed an estimator based on the finding that

$$E\{\frac{\delta_i Z_i}{G(Z_i)}|X_i\} = E(T_i|X_i) = X_i^T\beta \tag{10}$$

where $G(t) = Pr(C > t)$ is the survival function of the censoring time. But usually G is unknown and a natural thing to do is to replace G by an estimator in this quantity. Koul et al. (1981) proposed to estimate $G(Y_i)$ by

$$\hat{G}(t) = \prod_{J=1}^{n} \Big\{\frac{1 + N^+(Z_J)}{2 + N^+(Z_J)}\Big\}^{1_{\{\delta_J=0, Z_J\leq t\}}} \quad -\infty < t < +\infty \tag{11}$$

where $N^+(z)$ is the number of Z_I exceeding z, i.e.

$$N^+(z) = \sum 1_{\{Z_I>z\}} \tag{12}$$

The final weights are:

$$w_i = \frac{\delta_i}{\hat{G}(Z_i)} \tag{13}$$

Asymptotically, $\hat{G}$ behaves like the product-limit estimator of G. The inverse-probability weighted uncensored event time $\delta_i Z_i/\hat{G}(Z_i)$ or $w_i Z_i$ are regressed on X_i to estimate β. Another approach to estimate the left-hand limit of the Kaplan-Meier estimator of the censoring distribution $G(t) = Pr(C_i < t)$ is to use

$$\hat{G}(Z_i) = \prod_{i:Z_i\leq z} \Big[\frac{R_i}{B_i + R_i}\Big]^{I_{\{\delta_i=1\}}} \tag{14}$$

where R_j is the number of subjects "at risk" at time Z_j and B_j is the number of consecutive censored subjects before time Z_j (Leurgans, 1987). Thus, the weight function can be expressed as,

$$w_i = \frac{\delta_i}{\hat{G}(Z_i)} = \prod_{i:Z_i\leq z} \Big[\frac{B_i + R_i}{R_i}\Big]^{I_{\{\delta_i=1\}}} = \prod_{i:Z_i\leq z} \Big[\frac{R_{i-1} - D_{i-1}}{R_i}\Big]^{I_{\{\delta_i=1\}}} \tag{15}$$

$$R_i - D_i = B_i + R_i \tag{16}$$

where D_i is the number of failure subjects at time Z_i and $R_0 = D_0 = 0$.

This weight can be obtained by dividing the data into several more homogeneous groups and redistributing the weights of the censored observation according to $w_i = \frac{\delta_i}{\hat{G}(Z_i)}$ and $w_{(n)} = \frac{1}{\hat{G}(Z_{(n)})}$ only within the groups (Zhou, 1992).

Lemma 1.

$$w_i = n \times \textit{jump size of the Kaplan-Meier estimator of the } Z_i's = \frac{\delta_i}{\hat{G}(Z_i)}$$

Proof:

$$w_i = \tfrac{\delta_i}{\hat{G}(Z_i)} = \textstyle\prod_{i:Z_i\leq z} \Big[\frac{B_i+R_i}{R_i}\Big]^{I_{\{\delta_i=1\}}} = \prod_{i:Z_i\leq z} \Big[\frac{R_{i-1}-D_{i-1}}{R_i}\Big]^{I_{\{\delta_i=1\}}}$$
$$= \frac{R_0 - D_0}{R_1} \times \frac{R_1 - D_1}{R_2} \times ... \times \frac{R_{n-2} - D_{n-2}}{R_{n-1}} \times \frac{R_{n-1} - D_{n-1}}{R_n}$$

$$\Rightarrow \frac{R_0 - D_0}{R_1} \times \frac{R_1 - D_1}{R_2} \times ... \times \frac{R_{n-2} - D_{n-2}}{R_{n-1}} \times (R_{n-1} - D_{n-1}) = w_i R_n$$

$$n\Big[\hat{S}_{KM}(Z_{i-1}) - \hat{S}_{KM}(Z_i)\Big] = n \times \frac{1}{R_0} \times \frac{R_0 - D_0}{R_1} \times ... \times \frac{R_{n-2} - D_{n-2}}{R_{n-1}} \times (R_{n-1} - D_{n-1}) \times \Big[1 - \frac{R_n - D_n}{R_n}\Big]^{I_{\{\delta_n = 1\}}}$$

$$= n \times \frac{1}{n} \times w_i \times R_n \times \Big[\frac{D_n}{R_n}\Big] = w_i$$

assuming $D_n = 1$.

As a continuation to Koul et al. (1981)'s work, Leurgans (1987) proposed the "synthetic" data method, which transform the original data to synthetic data, by ranking Y_i, $i = 1, ..., n$ as $Y_{(1)}, Y_{(2)}, ..., Y_{(n)}$. The expression of the synthetic times is:

$$A_{(i)} = Z_{(1)} + \sum_{j=2}^{i} \frac{Z_{(j)} - Z_{(j-1)}}{\hat{G}(Z_{(j)})} = \int \frac{I(s < Z_{(i)})}{\hat{G}(Z_{(i)})} ds \tag{17}$$

β is estimated by regular least squares method where Y_i's were replaced by the synthetic times $A_{(i)}$.

$$\hat{\beta} = \arg\min_b \sum_{i=1}^{n} (A_{(i)} - X_{(i)}^T \beta)^2 \tag{18}$$

4 Inverse Probability Weighted Method

Similarly, Zhou (1992) proposed that, for any reasonable loss function $\rho(.)$, β can be estimated by minimizing an inverse probability weighted objective function. The weighted data are formed by properly redistributing the weights of the censored observations to the uncensored observations. For the last observation, Zhou (1992) modify the definition of the weight so that no weights will be lost in the redistribution process even if the largest observation is censored:

$$w_{(n)} = \frac{1}{\hat{G}(Z_{(n)})} \tag{19}$$

where $Z_{(n)} = max\{T_i\}$ denotes the largest observation and $w_{(n)}$ is the corresponding weight.

After the weighted data are formed, the estimator $\hat{\beta}$ of β is obtained by minimizing

$$f(b) = \sum_i \rho(Y_i - X_i^T b) w_i \tag{20}$$

with respect to b. Here ρ can be any meaningful loss function. For example, $\rho(y) = y^2$ for least squares, $\rho(y) = |y|$ for least absolute deviation. For example, minimizing $f(b) = \sum_i \rho(Y_i - X_i b) w_i$ with $\rho(y) = y^2$ and the simple linear model $E(Y_i) = \alpha + \beta x_i$ results in

$$\hat{\beta} = \frac{\sum [X_i - \bar{X}_w] w_i Y_i}{\sum [X_i - \bar{X}_w] w_i X_i} \tag{21}$$

where $\bar{X}_w = \sum w_i X_i / \sum w_i$.

A major advantage of this approach over the Buckley-James is that it does not need iteration in computing the estimator.

5 Simulation

The Standford heart transplantation data was used by many authors and can be found from the research of Miller and Halpern (1982). This program began in October 1967. By February 1980, 184 patients had received heart transplants. A few of these had multiple operations. Their survival times, uncensored or censored in February 1980, are included in the data along with their ages at the time of the first transplant. The patients' T5 mismatch scores which measure the degree of tissue incompatability between the initial donor and recipient hearts with respect to HLA antigens are also included in the data. The data on patients who were admitted to the program but did not receive transplants and the pretransplant waiting times for patients who did receive transplants were not collected for this study. Because of this there is no analysis of whether or not transplantation prolongs survival.

To assess the three Semi-Parametric models, simulation is performed based on the following model.

$$Y_i = 1.3 + 0.15Age - 0.0016Age^2 + \epsilon_i, \quad i = 1, ..., n \tag{22}$$

where $Y_i = log(T_i)$ and $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

The algorithm to perform the simulation can be briefly summarize as follows:

1. Using one sampling method to generate 300 errors from a designated distribution.
2. Randomly generate 300 observations (Y_is) by randomly sampling age from 12 to 64 and using the errors generated from step 1.
3. Randomly generate 300 indicators (200 uncensored and 100 censored) for the 300 observations.
4. Generate the inversed probability weights and synthetic data for each observation.
5. Compute the initial values using least-square methods on only uncensored data.
6. Find the estimators by minimize the objective functions.
7. Repeat Step 1-6 500 times and compute the average values for each parameter.

5.1 Inversion Method and Cauchy Distribution

When the error terms follow the Cauchy distribution with the location parameter = 0 and scale parameter = 1, the probability density function is

$$f_\epsilon(\epsilon) = \frac{1}{\pi(1+\epsilon)}, -\infty < \epsilon < \infty \tag{23}$$

To use the inversion method to simulate the error terms, we first find its cumulative distribution function:

$$
\begin{aligned}
F_\epsilon(\epsilon) &= \int_{-\infty}^{\epsilon} f_\epsilon(t)dt \\
&= \int_{-\infty}^{\epsilon} \frac{1}{\pi(1+t^2)}dt \\
&= \frac{1}{\pi}\int_{-\infty}^{\epsilon} \frac{1}{1+t^2}dt \\
&= \frac{1}{\pi}\Big[arctan(t)\Big]_{-\infty}^{\epsilon} \\
&= \frac{1}{\pi}\Big[arctan(\epsilon)+\frac{\pi}{2}\Big]
\end{aligned}
$$

$$F_\epsilon(\epsilon) = \frac{1}{\pi}arctan(\epsilon) + \frac{1}{2} = y \tag{24}$$

Then we calculate the inverse function:

$$
\begin{aligned}
y &= \frac{1}{\pi}arctan(\epsilon) + \frac{1}{2} \\
\Rightarrow (y - \frac{1}{2})\pi &= arctan(\epsilon) \\
\Rightarrow \epsilon &= tan((y-\frac{1}{2})\pi)
\end{aligned}
$$

$$F^{-1}(U) = tan((u-\frac{1}{2})\pi) \tag{25}$$

where $U \sim U[0,1]$.

Table 1: Simulation Results (Cauchy Distribution)

Model	*Intercept*	*Age*	Age^2
Synthetic	1.3293	0.1494	-0.0016
Least-squares (M-estimator)	1.2912	0.1067	-0.0017
LAD (M-estimator)	1.2917	0.1503	-0.0027

5.2 Rejection Method and t Distribution

When the error terms follow the t distribution degrees of freedom = 12, the probability density function is

$$f_\epsilon(\epsilon) = \frac{\Gamma(\frac{13}{2})}{\sqrt{12\pi}\Gamma(6)}\Big(1+\frac{\epsilon^2}{12}\Big)^{-\frac{13}{2}} \tag{26}$$

The rectangular rejection method was used to sample the error terms from the t distribution. The algorithm is as follows:

Taking $-10 \le \epsilon \le 10$, and $0 \le f_\epsilon(\epsilon) \le \theta$.

1. Generate $U \sim U[-10, 10]$.
2. Generate a second uniform random variable $V \sim V[0, 1]$ and set $Y = \theta V$
3. If $Y > f_\epsilon(U)$, we go back to step 1 and repeat, otherwise we accept the random variable U.

Table 2: Simulation Results (t Distribution, $\nu = 12$)

Type	$Intercept$	Age	Age^2
Synthetic	1.9495	0.1076	-0.0010
Least-squares Method	1.8586	0.0815	-0.0014
Least Absolute Deviation	1.8815	0.1120	-0.0021
Type-II	1.4596	0.1541	-0.0016

References

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

[2] L. Huang & Z. Jin. LSS: An S-Plus/R program for the accelerated failure time model to right censored data based on least-squares principle, *Computer Methods and Programs in Biomedicine*, 86 (2007) 45-50

[3] H. Koul, V. Susarla, & J. Van Ryzin. Regression Analysis with Randomly Right- Censored Data, *Ann. Statist*, 9 (1981) 1276-1288

[4] R.G. Miller & J. Halpern. Regression with censored data, *Biometrika*, 69 (1982) 521-531

[5] L. Pang. Semiparametric Estimation and Inference for Censored Regression Models, *Ph. D. thesis, North Carolina State University, North Carolina, USA* (2012)

[6] P.J. Rousseeuw, C. Croux, Alternatives to the median absolute deviation, *J. Amer. Statist. Assoc*, 88 (1993) 1273-1283

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

[8] M. Zhou, M-estimation in censored linear models, *Biometrika*, 79, (1992) 837-841