A Comparison of Two Methods for Estimating Censored Linear Regression Models

Ersin Yılmaz[*] and Dursun Aydın

Mugla Sitki Kocman University, Turkey

Abstract: This paper presents two basic methods called as weighted least squares (WLS) and synthetic data transformations (SDT). The key idea of the paper is to estimate the parameters of the linear regression model with randomly right-censored data by using these two methods. Recently, the mentioned methods have received considerable attention in the literature. Studies on this subject show that both methods work well for linear regression model with censored data. A particular focus of our paper is to compare the performance of the WLS and SDT methods and to reveal the strong and weak aspects of them. In this context, we made a simulation study and a real data example.

Keywords: Least squares, Kaplan-Meier estimator, linear regression, right-censored data, synthetic data, Kaplan-Meier weights.

1. INTRODUCTION

Consider the linear regression model

$$Y_i = X_{ij}\beta_i + \varepsilon_i,\ i=1,\ldots,n, j=1,\ldots,p \tag{1}$$

where X_i's are the values of realized covariate which are fully observed, Y_i's are the values of response variable, β is a $(p\times 1)$ parameter vector to be estimated, and ε_i's are independent and identically distributed with mean zero and constant variance.

In matrix and vector form, the model (1) is given by

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots X_{p1} \\ \vdots & \vdots & \ddots\ \vdots \\ 1 & X_{1n} & \cdots X_{pn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Then, the matrix and vector form of linear regression model (1) can be rewritten as

$$Y = X\beta + \varepsilon \tag{2}$$

In practice, Y_i's may be incompletely observed and the right censored by a censoring variable *C.* In this case, instead of observing (Y_i, X_i), we observe the data sets $\{(Y_i, X_i, \delta_i),\ i=1,\ldots,n\}$ with

$$T_i = \min(Y_i, C_i), \delta_i = I(Y_i \le C_i) = \{1\ if\ (Y_i \le C_i)\ and\ \ 0\ otherwise\} \tag{3}$$

where $I(.)$ is an indicator function that contains the information of censoring, T_i's are the observed lifetimes and C_i's are the censoring time, respectively, for the i th subject. In the presence of censoring, model (1) rewritten with updated response variables as

$$T_i = X_{ij}\beta_j + \varepsilon_i,\ i=1,\ldots,n,\ j=1,\ldots,p \tag{4}$$

It cannot be applied the ordinary least squares method for estimating the model (2) because of the variable T includes censored observations. To overcome this problem, there are two popular methods such as synthetic data transformations (SDT) and weighted least squares (WLS) with Kaplan-Meier weights. In this context, several important studies can be ordered as follows; Buckley and James [2] and Koul *et al.* [3] proposed the synthetic data transformation for modeling the censored data and they provide that original response variable *Y* and produced synthetic data have same expected values and asymptotic properties. Zheng [4] and Leurgans [5] gave extended properties of these synthetic data transformations and they compared mentioned two data transformation methods. Also, Zhou [6] and Lai *et al.* [7] studied about the asymptotic normality of synthetic data methods for regression with some applications. There are some other studies about the estimating the right-censored data with using synthetic data such as; Miller [8], Ritov [9], Tsiatis [10], Srinavasan [11], Fygenson [12], Li and Van Keilegom [13], Li and Wang [14], Wang and Dinse [15] respectively.

There are some specific studies about estimating the right-censored data with Kaplan-Meier weights; Miller [1] proposed the Kaplan-Meier weights and he estimated linear regression model with weighted least squares method. Stute ([16-18]) extended this method for nonlinear models and illustrated its consistency and asymptotic properties. Also Yu *et al.,* [19] used

[*] Address correspondence to this author at the Mugla Sitki Kocman University, Turkey; Tel: 905364251668;
E-mail: yilmazersin13@hotmail.com

weighted least squares method for estimating the censored data and they inspected the regression models for homoscedastic and heteroscedastic data. Khan and Shaw [20] studied about the weighted least squares method and in general, they deal with the treating the biggest censored value as an uncensored observation and they proposed some alternative adjusted imputation methods for estimation procedure.

In this paper, we focused on comparison of the mentioned two methods. The purpose of this study is to detect the advantages and disadvantages of SDT, and WLS based on Kaplan Meier weights [21] and we planned to uncover the superiorities of the methods according to each other. To realize our goal, we carried out a real data and a simulation study. To the best of our knowledge, such a study has not yet been made.

The paper is organized as follows. In Section 2, the data transformation method and Kaplan-Meier weights are expressed, respectively. The variances of the estimators are also illustrated in this section. In Section 3, simulation study and real data application are presented. Finally, conclusions and recommendations about the application are presented in the last section.

2. ESTIMATION OF LINEAR MODEL WITH RIGHT-CENSORED DATA

Let F, G and M be the distributions of the Y_i, C_i and T_i variables, respectively. According to these, survival functions of the mentioned variables could be written as

$$1-F\left(t/X\right)=P\left(Y_i>t/X\right),1-G\left(t/X\right)=P\left(C_i>t/X\right)$$

and because of the independence of Y_i and C_i

$$1-M(t/X)=(1-(F\,(t/X)\times G\,(t/X))=P(T_i>t/X).$$

In estimation procedure of censored data, it can be said that there are also two important assumptions

I. $\left(X_i, Y_i\right)$ and C_i are independent

II. $P\left(Y_i \le C_i / X_i, Y_i\right)=P\left(Y_i \le C_i / Y_i\right)$

Because of censoring, the conventional methods to estimate the right-censored response variable are unusable. As we said before, this problem arises due to T_i and true response variable Y_i have different expected values. In following two sections, we introduced two methods that overcome this censorship problem.

2.1. Synthetic Data Transformation

The SDT method is proposed by [3] to overcome the problems caused by censored data. In summary, the SDT provides the equality, $E(T_i)=E(Y_i)$ and by some modifications on censored and uncensored observations. Where $E(.)$ denotes the expected value (see, [16]). In our context, data transformation could be given by

$$T_{iG}=\frac{\delta_i T_i}{1-G\left(T_i\right)} \tag{5}$$

Where G (.) is the distribution of the censoring variable C_i, as expressed in the section 2. Thus, model (1) can be rewritten as

$$T_{iG}=X_{ij}\beta_j+\varepsilon_{iG}, \quad 1\le i\le n, 1\le j\le p \tag{6}$$

Where the ε_{iG}'s are the error terms for a known G. Generally, distribution G is unknown and need to be estimated. To solve this problem, [3] used Kaplan-Meier estimator defined by

$$1-\hat{G}\left(t\right)=\prod_{i=1}^{n}\left(\frac{n-i}{n-i+1}\right)^{I\left[T_{(i)}\le t,\delta_{(i)}=0\right]},(t\ge 0) \tag{7}$$

where $T_{(i)}$'s are the ordered values of censored response variable such as $T_{(1)}\le T_{(2)}\le\cdots\le T_{(n)}$ and $\delta_{(i)}$'s are the ordered indicator values associating with $T_{(i)}$'s. Furthermore, $\hat{G}\left(t\right)$ has jumps only at the censored observations (see, [22]).

In order to estimate the parameter of the model (6), ordinary least square (OLS) method is used with replacing censored response variable with synthetic response variable T_{iG}. Thus, the estimation of the regression coefficients β_{ij}'s could be obtained by solving the minimization criterion

$$RSS\left(\beta_j\right)=\sum\nolimits_{i=1}^{n}\left(T_{iG}-X_{ij}\,\hat{\beta}_j\right)^2 \tag{8}$$

where T_{iG}'s are unknown values of synthetic response variable. It should be noted they are estimated by $T_{i\hat{G}}=\delta_i T_i/\left(1-\hat{G}\left(T_i\right)\right)$. Hence, the equation (8) is updated as follows

$$RSS\left(\beta_j\right)=\sum\nolimits_{i=1}^{n}\left(T_{i\hat{G}}-X_{ij}\,\hat{\beta}_{ij}\right)^2=\left(T_{i\hat{G}}-X\hat{\beta}_s\right)'\left(T_{i\hat{G}}-X\hat{\beta}_s\right) \tag{9}$$

where $\hat{\beta}_s$ represents the estimated coefficient vector obtained by synthetic data.Some algebraic calculations show that the estimated vector β_s that minimizes the criterion (9) is calculated as

$$\hat{\beta}_s = \left(X'X\right)^{-1} X'T_{i\hat{G}} \tag{10}$$

Note also that it can be said that $\hat{\beta}_s$ is a biased estimator of the**,** because of the synthetic data. However, it is heuristically said that when the sample size is getting infinity, bias term converges to zero. From the equation (9) the variance of the model is obtained as

$$\hat{\sigma}_S^2 = RSS\left(\hat{\beta}_S\right)/\left(n-p\right) \tag{11}$$

where $\left(n-p\right)$ is the degrees of freedom. Also, covariance matrix of the estimator is defined as follows

$$Var\left(\hat{\beta}_S\right) = \hat{\sigma}_S^2 \left(X'X\right)^{-1} \tag{12}$$

The diagonal elements of this matrix denote the variances of the estimators $\left(\hat{\beta}_i, i=1,\ldots,p\right)$ of the individual parameters, while the off-diagonal elements indicate the co-variances among these estimators.

2.2. Weighted Least Squares with Kaplan-Meier Weights

The WLS estimator of $\hat{\beta}_W$ is defined by

$$\hat{\beta}_W = \arg\min\left[\sum_{i=1}^{n} W_i \left(T_i - X_i\beta\right)^2\right], \tag{13}$$

where W_i's are the Kaplan-Meier weights obtained by Kaplan-Meier estimator. Calculations of weights are based on the jumps of the K-M estimator. According to (3) Kaplan-Meier weights are obtained as follows

$$W_i = \frac{\delta_{(i)}}{n-i+1}\prod_{j=1}^{i-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}}, \quad i=2,\ldots,n \;\; and \; W_1 = \frac{\delta_1}{n} \tag{14}$$

where $T_{(i)}$ is the ith minimum value of T and $\delta_{(i)}$'s are the ordered values associated with $T_{(i)}$'s. It can be seen from (14), this weight function gives zero weight to censored observations and the largest observation $T_{(n)}$. Thus, from equation (14) weighted least squares estimate of the $\hat{\beta}_W$ is obtained as

$$\hat{\beta}_W = \left(X'WX\right)^{-1} X'WT \tag{15}$$

Where X is the $n\,x\,p$ matrix, W is the $n\,x\,n$ dimensional weight matrix, and T is the right-censored response vector.

As in the equation (12), the variance of the estimator $\hat{\beta}_W$ is calculated by using the sum of squares of residuals

$$RSS(\beta) = \left(T_i - X\hat{\beta}_W\right)' W \left(T_i - X\hat{\beta}_W\right)$$

Hence, the estimator of variance of the model (2) is

$$\hat{\sigma}_W^2 = \left(T_i - X\hat{\beta}_W\right)' W \left(T_i - X\hat{\beta}_W\right)/\left(n-p\right) \tag{16}$$

and the variance-covariance matrix of the $\hat{\beta}_W$ can be obtained as follows is

$$Var\left(\hat{\beta}_W\right) = \hat{\sigma}_W^2 \left(X'WX\right)^{-1} \tag{17}$$

To gain some understanding of how well the mentioned methods work, we obtained means and standard errors for the estimates obtained by the WLS and the SDT methods under the three different censoring levels. Moreover, in order to assess the quality of the regression parameters, we used Mean Squared Errors (MSEs) and coefficient of determination which can be calculated as, respectively;

$$MSE = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(T_i - \hat{T}_l\right)^2 \tag{18}$$

In this study, because we are dealing with the linear regression models, we can use the coefficient of determination R^2 for measuring the quality of estimations. R^2 can be defined by

$$R^2 = \frac{\sum\nolimits_{i=1}^{n}\left(\hat{T}_l - \bar{T}_l\right)^2}{\sum\nolimits_{i=1}^{n}\left(T_i - \bar{T}_l\right)^2} \tag{19}$$

where T_i's are the fitted values $\bar{T}$, is the mean values of response variable T .

3. NUMERICAL EXAMPLES

3.1. Simulation Study

For convenience, we carried out a simulation study to compare the performance of the two methods stated in the previous section. There are 1000 simulation runs for three different sample sizes (n=50,100, 250) and censoring levels (C.L.=10%, 25%, 40%). The outcome T is generated according to a censored linear regression model

$$T_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_i, i=1,\ldots,n \tag{20}$$

where $X_i \sim N(0,1)$, $\beta_i = \left(\alpha,\beta_1,\beta_2,\beta_3\right)^T = \left(1,3,-2,\ 0.5\right)^T$ and ε_i's are the random error terms from $N(0,1)$. Here, T_i is the randomly-right censored response variable which is obtained as (3). We generate the censoring variable C_i from Bernoulli distribution for three different probabilities. The following figures and tables summarize the results that obtained from model (20).

When we inspected the Figure **1**, as the sample sizes increase, the range of the estimates becomes narrower as expected. However, it can be clearly seen that the estimates from the WLS method are better than those of the SDT method under the lower censoring levels.

The outcomes in the Figure **2** show that the ranges of the estimates are wider because of the high censoring level. When we evaluate the Table **1** and the Figure **2** together, the key point is that the SDT method corrupted more than WLS. Therefore, it can be said that the WLS is affected from censorship level change less than the SDT method for this simulation study.

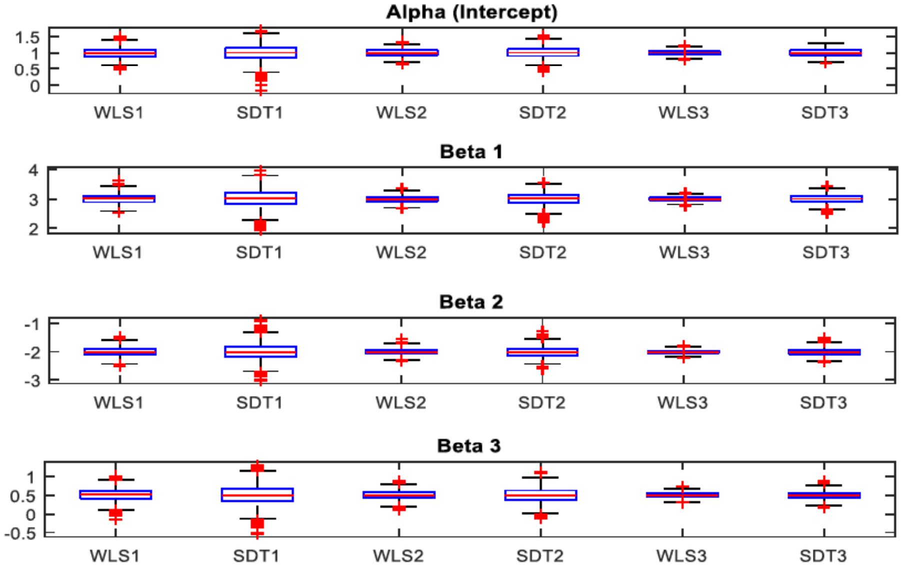

Figure 1: In x-axis of each boxplot "WLS1", "WLS2" and "WLS3" represent the estimated regression coefficients obtained by KM weights for sample sizes 50, 100 and 250, respectively, and C.L.=10%."SDT1", "SDT2" and "SDT3" are similar to WLS1", "WLS2" and "WLS3, respectively, but for SDT method.

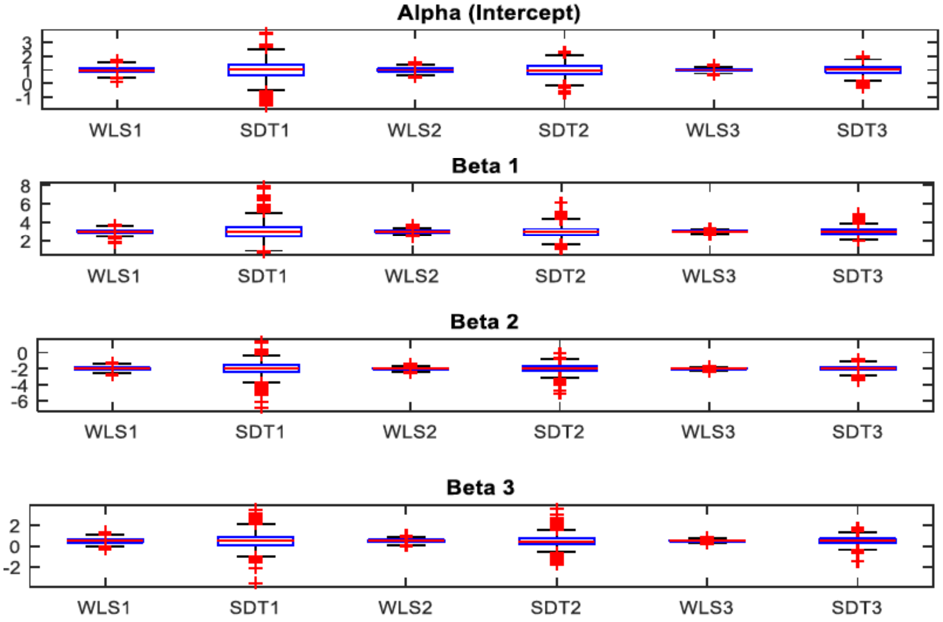

Figure 2: Similar to Figure 1 but for C.L.=40%.

Table 1: **Estimates and Standard Deviations of the Regression Coefficients**

		WLS				SDT			
		$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	50	0,9920	3,0045	-2,0018	0,5119	0,9863	3,0242	-2,0077	0,5076
		(0,1425)	**(0,1225)**	(0,1388)	(0,1665)	(0,1541)	(0,1325)	(0,1501)	(0,1801)
	100	0,9986	2,9985	-1,9980	0,5021	1,0035	3,0082	-2,0064	0,5038
		(0,0992)	(0,1014)	**(0,0904)**	(0,0953)	(0,1027)	(0,1050)	(0,0936)	(0,0986)
	250	1,0029	2,9994	-2,0001	0,5015	1,0021	3,0033	-2,0043	0,4977
		(0,0632)	(0,0659)	(0,0656)	**(0,0612)**	(0,0641)	(0,0668)	(0,0666)	(0,0621)
25%	50	0,9949	3,0023	-1,992	0,4914	1,0031	3,0283	-2,0232	0,4856
		(0,1397)	(0,1424)	(0,1659)	(0,1515)	**(0,1170)**	(0,1804)	(0,2102)	(0,1919)
	100	1,0002	3,0024	-1,9984	0,4976	1,0117	3,0266	-2,0092	0,5071
		(0,1007)	(0,1091)	(0,1093)	(0,1093)	(0,1152)	(0,1247)	(0,1249)	(0,1250)
	250	1,0018	3,0052	-1,9964	0,4957	1,0130	3,0117	-2,0090	0,4953
		(0,0364)	(0,0631)	(0,0641)	(0,0654)	(0,067)	(0,0668)	(0,0677)	(0,0692)
40%	50	0,9943	3,0053	-2,0114	0,4957	0,9755	3,0169	-2,0134	0,5063
		(0,1459)	(0,1667)	**(0,1373)**	(0,1551)	(0,2303)	(0,2631)	(0,2167)	(0,2449)
	100	0,9913	3,0025	-1,9980	0,4911	0,9792	2,9831	-1,9890	0,4621
		(0,1036)	(0,1036)	**(0,1033)**	(0,1035)	(0,1367)	(0,1367)	(0,1363)	(0,1366)
	250	0,9985	2,9977	-1,9998	0,5017	0,9993	3,0068	-1,9912	0,5002
		(0,0636)	(0,0642)	(0,0661)	(0,0639)	(0,0732)	(0,0740)	(0,0762)	(0,0737)

Table **1** shows the mean values of regression coefficients and standard deviations (in parentheses) obtained from 1000 simulated data sets. The results in the Table **1** indicate that the magnitudes of variances are decreasing as sample sizes are getting larger. As expected, the quality of estimates deteriorates under high censoring levels, in details, when we look at the results of two estimation methods; we observed that WLS method gave more satisfying results than SDT according to variances. In Table **1**, minimum variances values are denoted by bold color and here, almost all of the minimum variances are obtained by WLS method.

The overall point is that the performances of the two methods are quite different, especially in higher censoring levels. The outcomes from the data under the C.L=40% prove that WLS has improved the performances over SDT. In order to be sure about the superiority of the WLS method we present the plot of the variances of the regression models for different sample sizes and censoring levels in Figure **3**. In this figure, left panel represents the variances for C.L. = 10% and similarly right panel is designed for the C.L. = 40%.

When we examine the Figure **3**, we can clearly said that the linear regression models estimated by WLS method have smaller variances than variances of models that obtained by SDT for all sample sizes and all censoring levels. Furthermore, when censoring level changes from low to high, we cannot see an excessive increment on variances of WLS, we measure the changing as about 0.03, but in SDT method, variances change from 1.1 to 3 (difference is 2.9) which means that censoring level effects the estimations of SDT method badly.

To see more specifically the difference between qualities of the two methods Table **2** is given below that includes the MSE values and scores of R^2. It is well known that lower MSE and bigger values indicate better model fit to the data. For these purposes, the MSE and values from the methods are calculated and they are summarized in Table **2**. These outcomes show that when the censoring levels are 25% and 40%, the WLS fits outperform the SDT fits for all sample sizes. This indicates that the WLS method is preferred when the data censoring rate is moderate and high.

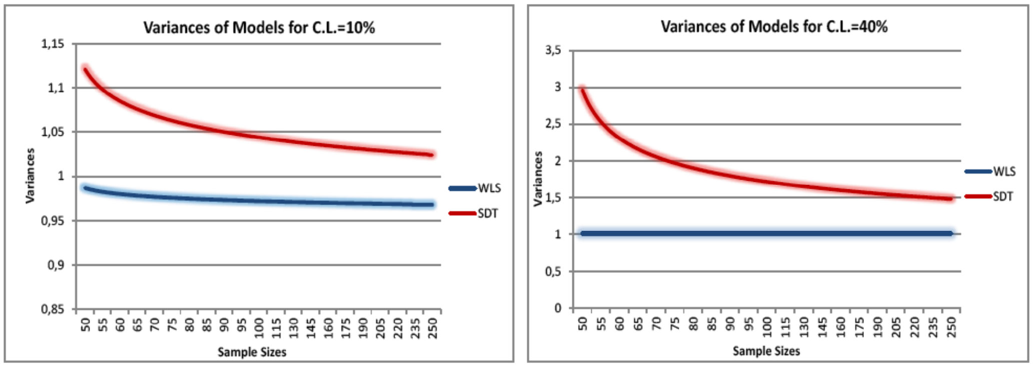

Figure 3: Distributions of the variances obtained from WLS and SDT methods for different sample size.

Table 2: MSE and Values for Comparing the WLS Fits of Model (20) Against the Associated SDT Fits

n=50						
	MSE			R^2		
C.L.	10%	25%	40%	10%	25%	40%
WLS	0,0009	0,0008	0,0009	0,931	0,9283	0,9244
SDT	0,0014	0,0021	0,0026	0,9167	0,8763	0,7987
n=100						
	MSE			R^2		
C.L.	10%	25%	40%	10%	25%	40%
WLS	0,0012	0,0007	0,0009	0,9297	0,9287	0,9272
SDT	0,0012	0,001	0,0025	0,9231	0,903	0,8536
n=250						
	MSE			R^2		
C.L.	10%	25%	40%	10%	25%	40%
WLS	0,0009	0,0008	0,0011	0,9298	0,9296	0,9284
SDT	0,0009	0,0009	0,0017	0,9273	0,9192	0,8968

However, it should be noted that if the data is uncensored, any of these methods may be preferred.

3.2. Real Data Example

In this section, we presented the results of real data application. We made the experiment with data collected from colon cancer patients in Izmir, Turkey. In this example, we used the logarithm of the survival times of patients as a response variable $(Stime)$. Eight independent variables are denoted as: *sex*, *age*, application (app), location of the tumor (loc), score of organ and tissue transplant (tx), liver metastasis (met), type of operation (op) and phase of cancer $(phase)$. For these variables, we will fit a linear regression model given by

$$\log(Stime_i) = \beta_0 + \beta_1 sex_i + \beta_2 age_i + \beta_3 app_i + \beta_4 loc_i + \beta_5 tx_i + \beta_6 met_i + \beta_7 op_i + \beta_8 phase_i + \varepsilon_i \tag{21}$$

where $i = 1,\ldots,40$. A total of 40 patients are observed in the analysis. Of the 40 patients in the sample, 11 are censored from the right randomly for different reasons such as withdrawing from study or death from different illness. So, the censorship rate is calculated as 27.50%.

Regression results obtained by two methods are illustrated in table and figure. Table **3** shows the estimates and variances of regression coefficients, and the values of the MSE and the coefficient of determination R^2 for the mentioned two method.

Table 3: Comparative Outcomes for the WLS and SDT Methods

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	MSE	R^2
WLS	3,9593	-0,5397	0,0230	0,5976	-0,0434	-0,0665	-0,3753	0,0527	-0,3760	**0,990**	**0,344**
	(1,991)	(0,123)	(0,0004)	(0,153)	(0,013)	(0,002)	(0,138)	(0,004)	(0,046)		
SDT	0,8459	-0,3281	0,0259	1,2784	-0,3568	-0,1144	-2,3537	0,2159	-0,1439	3,914	0,321
	(7,8806)	(0,482)	(0,002)	(0,606)	(0,053)	(0,007)	(0,546)	(0,016)	(0,181)		

As expressed earlier, the outcomes from the WLS and SDT are compared in Table **3**. It should be noted that the variances of the estimated regression coefficients are given in parentheses. From Table **3**, we observe that the variability measures of the estimates, the values of MSE and R^2 obtained by the WLS are smaller than those of the SDT. So, the WLS method giving the smallest values is preferred.

Note that some values of the response variables are censored. In this context, Figure **4** shows the real survival time data with together fitted values from two methods, SDT and WLS, respectively. As can be seen in Figure **4**, fitted values from the WLS (indicated by ◘) are more stable and consistent than fitted values obtained from SDT (marked by×). One of the most important reasons for this is that a synthetic response variable whose expectation is equal to the original one and then gets the least squares estimator by using this unbiased synthetic response variable. As expected, the weighted least squares (WLS) method gives a better performance than the least squares using synthetic response variable (SDT) in this study. Furthermore, note also that when the sample size is getting larger, mentioned methods are beginning to give almost same results for estimating models.

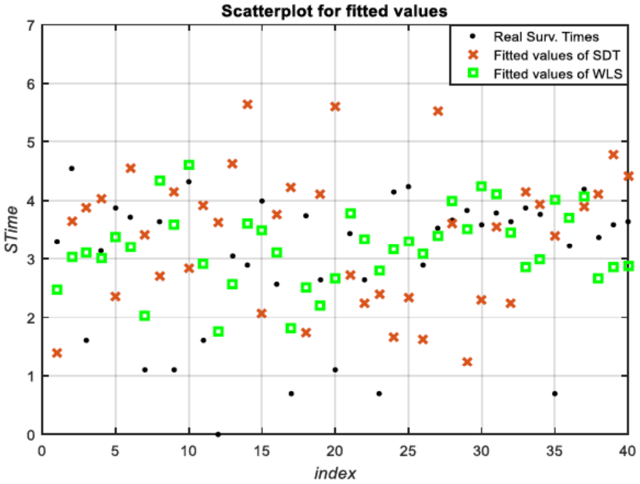

Figure 4: Scatter plot for the survival times of patients: Character "•" denotes the real survival times. Also, character " × " indicates the fitted values from the SDT, while the fitted values from the WLS are marked by "◘".

4. RECOMMENDATIONS AND CONCLUSIONS

In this study, we focus attention on estimation and comparison of the WLS and SDT methods. To realize these purposes we carried out a simulation study and a real data application. Then it is obtained some results about these two methods. Accordingly, we interpreted the results and listed our comments.

- For low censoring levels, the difference of estimation quality between two methods is almost negligible especially for large samples which can be seen in Figure **1** and Table **1**.
- In high censoring levels and small sample sizes WLS method can resist the censorship but performance of SDT begin to decrease.
- Thus, we are recommending the WLS method for small sample sizes, and high censoring levels but according to this study, SDT can be used for the large sample sizes or the low censoring levels because it gives good results under certain conditions.
- The WLS method is found to be better than SDT when these methods are applied to the data set containing 8 features of colon cancer patients.
- The results obtained by the real data sample and simulation study are consistent with each other.

REFERENCES

[1] Miller RG. Least Squares Regression with Censored Data. Biometrika 1976; 63: 449-464. https://doi.org/10.1093/biomet/63.3.449

[2] Buckley J, James I. Linear regression with censored data. Biometrika 1979; 66(3): 429-436. https://doi.org/10.1093/biomet/66.3.429

[3] Koul H, Susarla V, Van Ryzin J. Regression Analysis with Randomly Right-Censored Data. The Annals of Statistics

1981; 9(6): 1276-1285. https://doi.org/10.1214/aos/1176345644

[4] Zheng ZK. Regression Analysis with Censored Data. PhD Dissertation1984; University of Colombia.

[5] Leurgans S. Linear models, random censoring and synthetic data. Biometrika 1987; 74: 301-309. https://doi.org/10.2307/2336144

[6] Zhou M. Asymptotic Normality of the 'Synthetic Data' Regression Estimator for Censored Survival Data. The Annals of Statistics1992; 20(2): 1002-1021. https://doi.org/10.1214/aos/1176348667

[7] Lai TL, Ying Z and Zheng ZK. Asymptotic normality of a class of adaptive statistics with applications to synthetic data methods for censored regression. Journal of Multivariate analysis 1995; 52: 259-279. https://doi.org/10.1006/jmva.1995.1013

[8] Miller R, Halpern J. Regression with Censored Data. Biometrika, 1982; 69(3): 521-531. https://doi.org/10.1093/biomet/69.3.521

[9] Ritov Y. Estimation in a Linear Regression Model with Censored Data. The Annals of Statistics 1990; 18(1): 303-328. https://doi.org/10.1214/aos/1176347502

[10] Tsiatis AA. Estimating Regression Parameters Using Linear Rank Tests for Censored Data. The Annals of Statistics1990; 18(1): 354-372. https://doi.org/10.1214/aos/1176347504

[11] Srinavasan C and Zhou M. Linear Regression with Censoring. Journal of Multivariate Analysis 1994; 49: 179-201. https://doi.org/10.1006/jmva.1994.1021

[12] Fygenson M and Zhou M. On Using Stratification in the Analysis of Linear Regression Models with Right Censoring. The Annals of Statistics 1994; 23(2): 747-762. https://doi.org/10.1214/aos/1176325494

[13] Li G and Van Keilegom I. Likelihood Ratio Confidence Bands in Nonparametric Regression with Censored Data. Scandinavian Journal of Statistics 2002; 29(3): 547-562. https://doi.org/10.1111/1467-9469.00305

[14] Li G and Wang QH. Empirical Likelihood Regression Analysis for Right Censored Data. Statistica Sinica 2003; 13: 51-68.

[15] Wang Q and Dinse GE. Linear Regression Analysis of Survival Data with Missing Censoring Indicators. Lifetime Data Analysis2011; 17(2): 256-279. https://doi.org/10.1007/s10985-010-9175-8

[16] Stute W. Consistent Estimation under Random Censorship When Covariablesare Present. Journal of Multivariate Analysis 1993; 45: 89-103. https://doi.org/10.1006/jmva.1993.1028

[17] Stute W. The Central Limit Theorem under Random Censorship. The Annals of Statistics 1995; 2: 422-439. https://doi.org/10.1214/aos/1176324528

[18] Stute W. Nonlinear Censored Regression. Statistica Sinica 1999; 9: 1089-1102.

[19] Yu L, Liu L and Chen DG. Weighted Least-Squares Method for Right Censored Data in Accelerated Failure Time Model. Biometrics 2013; 69(2): 358-365. https://doi.org/10.1111/biom.12032

[20] Khan HR and Shaw JEH. Variable Selection for Survival Data with a class of Adaptive Elastic Net Techniques. Statistics and Computing 2016; 26(3): 725-741. https://doi.org/10.1007/s11222-015-9555-8

[21] Kaplan EL and Meier P. Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association 1958; 53(282): 457-481. https://doi.org/10.1080/01621459.1958.10501452

[22] Peterson AV Jr. Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions. Journal of the American Statistical Association 1977; 72: 854-858.

Received on 02-07-2017 Accepted on 07-08-2017 Published on 30-08-2017