International Journal of Advanced Statistics and Probability, *5 (1) (2017) 23-31*

International Journal of Advanced Statistics and Probability

Website: www.sciencepubco.com/index.php/IJASP
doi: 10.14419/ijasp.v5i1.7221
Research paper

Estimation and application in log-Fréchet regression model using censored data

Hanan H. Alamoudi [1, 3] *, Salwa A. Mousa [1, 2], Lamya A. Baharith [1]

[1] Department of Statistics, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia
[2] Faculty of Commerce (Girls Campus), Al-Azhar University, Nasr city, Cairo, Egypt
[3] Department of Statistics, Faculty of Science, King Abdulaziz University, P.O.Box 42805, Jeddah 21551, Saudi Arabia
**Corresponding author E-mail: hhalamoudi@kau.edu.sa*

Abstract

This article introduces a new location-scale regression model based on a log-Fréchet distribution. Maximum likelihood and Jackknife methods are used to estimate the new model parameters for censored data. Martingale and deviance residuals are obtained to check model assumptions, data validity, and detect outliers. Moreover, global influence is used to detect influential observations. Monte Carlo simulation study is provided to compare the performance of the maximum likelihood and jackknife estimators for different sample sizes and censoring percentages. The empirical distribution of the martingale and deviance residuals of the proposed model is examined. A real lifetime heart transplant data is analyzed under the log-Fréchet regression model to illustrate the satisfactory results of the proposed model.

Keywords: *Fréchet Distribution; Regression Model; Censored Data; Maximum Likelihood; Jackknife Method; Residual Analysis.*

1. Introduction

The study of the effect of covariate variables on survival time is crucial in many practical applications. Many regression models help to measure this effect. The log-location-scale model is considered as an important type of parametric regression model and commonly used in clinical trials, see Lawless [1]. This model assumes a linear relationship between the log of lifetime T and the covariate variables.

Several studies were conducted using the log-location-scale regression models. These include the log-Burr XII regression model with censored data analysis by Silva et al. [2], the log-modified Weibull regression models with censored data by Carrasco, Ortega and Paula [3], the log-generalized modified Weibull regression model with censored analysis by Ortega, Cordeiro and Carrasco [4], the log-exponentiated Weibull regression model with interval-censored analysis by Hashimoto et al. [5], the log-Weibull extended regression model with censored data analysis by Silva, Ortega and Cancho [6], the log-Burr XII regression model with grouped survival data analysis by Hashimoto et al. [7], log-odd log-logistic Weibull regression model with censored data by Cruz, Ortega and Cordeiro [8], log-odd log-logistic generalized half-normal regression model with censored data by Pescim et al. [9].

In this paper, based on log-location-scale regression model and Fréchet distribution, the log-Fréchet (LF) regression model is proposed. Fréchet distribution is considered as type II of the extreme value distribution.There are many applications for extreme value distributions in several fields such as floods, earthquakes, rainfall, sporting, wind speed, queues in supermarkets, and others, see Kotz and Nadarajah [10]. Moreover, the aspects of classical analysis for modeling censored data based on LF regression models are examined. Asymptotic distribution of the maximum likelihood (ML) estimators is carried out which is useful for small sample size since the normality assumptions is not easy to validate. Therefore, the use jackknife estimator is explored for the LF regression model for censored data. Moreover, it is important to examine the assumptions and performs a diagnosis approach after modeling in order to detect influential and outlying observations, see Cook [11].

The article is structured as follows:

Section 2 displays the log-Fréchet distribution. The log-Fréchet regression model is introduced in Section 3. Section 4 presents estimates of model parameters using maximum likelihood and jackknife methods based on censored data. Global sensitivity analysis is discussed in Section 5. In Section 6, residuals analysis is conducted to assess departures from the underlying log-Fréchet model and to detect outliers. Section 7 presents simulation study to estimate model parameters. In Section 8, a medical data set is investigated to show the flexibility and practically of the new regression model. Finally, concluding remarks are presented in Section 9.

2. The log-Fréchet distribution

This distribution was introduced by Maurice Fréchet (1878-1973). Assuming that the random variable T follows the standard Fréchet distribution with parameters λ and γ. The cumulative distribution function (CDF) and the corresponding probability density function (PDF) are respectively given by:

$$F(t;\lambda,\gamma) = exp\left\{-\left(\frac{\gamma}{t}\right)^{\lambda}\right\} \; t > 0,$$

$$f(t;\lambda,\gamma) = \lambda\gamma^{\lambda}t^{-(\lambda+1)}exp\left\{-\left(\frac{\gamma}{t}\right)^{\lambda}\right\}. \qquad (1)$$

Then the random variable $Y = log(T)$ will have log-Fréchet distribution (LFD) with transformation parameter $\sigma = 1/\lambda$ and $\mu =$

$log(\gamma)$. Therefore, the PDF and CDF for LFD are given as follows:

$$F(y;\sigma,\mu) = exp\left\{-exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\}, -\infty < y < \infty\,,$$

$$f(y;\sigma,\mu) = \frac{1}{\sigma}\, exp\left\{-\left(\frac{y-\mu}{\sigma}\right)\right\}\, exp\left\{-exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\}, \qquad (2)$$

where $\sigma > 0$ and $-\infty < \mu < \infty$ are the scale and location parameters respectively. The survival and hazard functions respectively are as follows:

$$S(y;\alpha,\sigma,\mu) = 1 - exp\left\{-exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\},$$

$$h(y;\alpha,\sigma,\mu) = \frac{exp\left\{-\left(\frac{y-\mu}{\sigma}\right)-exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\}}{\sigma\left[1-exp\left\{-exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\}\right]}.$$

Fig. 1 shows that the LFD has a single mode. The LFD has a monotonic increasing survival function and a monotonic decreasing hazard function.

For the variable Y, the moment generating function is:
$\mathrm{M_Y(t)} = \exp(\mathrm{t\mu})\ \Gamma(1-\mathrm{t\sigma}), \mathrm{t\sigma} < 1$

The standardized random variable $Z = (y-\mu)/\sigma$ has density function given by:
$$f(z) = exp\{-z\}\, exp\{-exp[-z]\}, -\infty < z < \infty. \qquad (3)$$

And its survival function is:
$S(z) = 1 - exp\{-exp[-z]\}$.

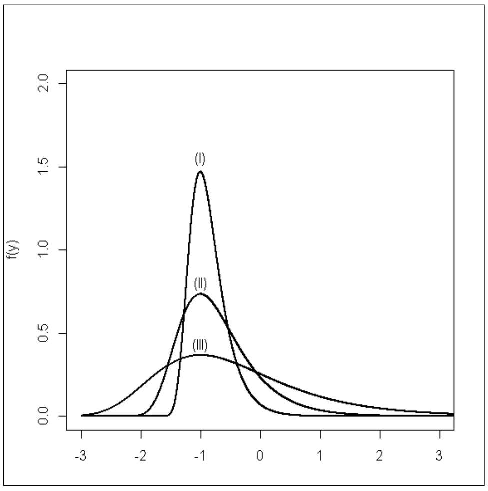


Fig. 1: PDF of the LFD with Parameter $\mu = -1$ and Different Values of Parameter $\sigma =$ (I) 0.25, (II) 0.5, And (III)1.

3. The log-Fréchet regression model

In many real applications for lifetime data, investigation of the relation between the survival time and the independent (explanatory) variables is important. Therefore, the regression model approach can be used. That is, the model in (2) can be written as linear log-location-scale regression model:

$$y_i = \beta^T x_i + \sigma z_i\ i = 1,2,3,\dots,n\,, \qquad (4)$$

where z_i is the random error with density function in (3), $\beta = (\beta_1,\dots,\beta_p)^T, \sigma > 0$ is a vector of unknown parameters, and $x_i = (x_{i1},\dots,x_{ip})^T$ is the explanatory variable vector. The parameter $\mu_i = \beta^T x_i$ is the location of Y_i. The location parameter vector $\mu = (\mu_1,\dots,\mu_n)^T$ can be represented as a linear model $\mu = \beta^T X$, where $X = (x_1,\dots,x_n)^T$ is a known model matrix. In this case, the survival function of Y|x is given by:

$$S(y|x) = 1 - exp\left\{-exp\left[-\left(\frac{y-\beta^T x_i}{\sigma}\right)\right]\right\}.$$

4. Estimation

4.1. Maximum likelihood estimation

Let $(y_1,\tau_1,x_1),\dots,(y_n,\tau_n,x_n)$ be a right censored random sample of n observation, where $y_i = \begin{cases} log(t_i)\ if\ \tau_i = 1 \\ log(c_i)\ if\ \tau_i = 0 \end{cases}, t_i$ and c_i are lifetimes and censoring times respectively and x_i is an explanatory variable. Assuming that the life times and censoring times are random and independent, the log likelihood function is given by:

$$l(\boldsymbol{\theta}) = -rlog(\sigma) + \textstyle\sum_{i=1}^{n}\tau_i[-z_i - exp(-z_i)] \\ + \textstyle\sum_{i=1}^{n}(1-\tau_i)log\big[1 - exp[-exp(-z_i)]\big], \qquad (5)$$

where r denotes the number of uncensored observations, $\boldsymbol{\theta} = (\sigma,\beta)^T$ and $z_i = (y_i - \beta^T x_i)/\sigma$.

By maximizing the log likelihood in (5), the maximum likelihood estimate (MLE) for the parameter vector $\boldsymbol{\theta}$ can be obtained. The *nlminb* function in the statistical package *R* is used to obtain the MLE. Also, the covariance estimates for $\widehat{\boldsymbol{\theta}}$ is acquired from the *Hessian* matrix. Under some regularity conditions, the distribution of $\widehat{\boldsymbol{\theta}}$ is asymptotically normal with covariance matrix that represents the inverse of the Fisher information matrix $I(\boldsymbol{\theta})^{-1}$ where $I(\boldsymbol{\theta}) = E\left[-\left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\right)\right]$. However, the presence of censored observations makes the computation of the Fisher information matrix difficult. Therefore, the second derivatives matrix of the log-likelihood can be derived and evaluated at the maximum likelihood estimator $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. Then the asymptotic normal approximation for $\widehat{\boldsymbol{\theta}}$ could be defined as $\widehat{\boldsymbol{\theta}}^T \sim N_{(p+1)}\{\boldsymbol{\theta}^T, \ddot{L}(\boldsymbol{\theta})^{-1}\}$, where $\ddot{L}(\boldsymbol{\theta}) = -\left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\right)$ is the $(p+1)\times(p+1)$ observed matrix such that:

$$\ddot{L}(\boldsymbol{\theta}) = \begin{pmatrix} -\left.\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\sigma^2}\right|_{\widehat{\beta}_j,\widehat{\sigma}} & -\left.\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\sigma\partial\beta_j}\right|_{\widehat{\beta}_j,\widehat{\sigma}} \\ -\left.\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\beta_j\partial\sigma}\right|_{\widehat{\beta}_j,\widehat{\sigma}} & -\left.\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\beta_j\partial\beta_s}\right|_{\widehat{\beta}_j,\widehat{\beta}_s,\widehat{\sigma}} \end{pmatrix},$$

where $i, j = 1,\dots,p$, and the submatrices are defined in Appendix A.

4.2. Jackknife estimation

Jackknife estimation is used to improve the estimate of the parameter by reducing the bias of the estimate. The method is based on "leave one out" procedure. Miller [12] uses jackknife in linear models to estimate the variance and bias of model parameters, see Tu and Shao [13]. The idea of jackknife method depends on transforming the problem of estimating population parameters into the problem of estimating the mean of the population.
Suppose that $T_1, T_2, \dots, T_n$ is a random sample and the sample mean $\bar{T} = \sum_{i=1}^{n}\frac{T_i}{n}$ is used to estimate population mean. Then the mean sample when lth observation deleted can be obtained as:

$$\bar{T}_{-l} = \frac{\sum_{i=1}^{n} T_i - T_l}{n-1} \text{ for which,}$$

$$T_l = n\hat{T} - (n-1)\bar{T}_{-l} \qquad (6)$$

Let $\hat{\theta}$ be the parameter estimator of the whole sample. Then $\hat{\theta}_{-l}$ is the parameter estimate when we drop the lth observation from the sample. The pseudo-value of the lth observation can be calculated from (6) as the difference between parameter estimation from whole sample and parameter estimation obtained without the lth observation. That is:
$\tilde{\theta}_l = n\hat{\theta} - (n-1)\hat{\theta}_{-l}$

The Jackknife estimate of θ, denoted by $\hat{\theta}_{jack}$ is the average of pseudo-values:

$$\hat{\theta}_{jack} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\theta}_i,$$

for more details, see Abdi and Williams [14]. Therefore, the jackknife bias estimator is:

$$b_{jack} = (n-1)(\hat{\theta}_{jack} - \hat{\theta})$$

An approximate $100(1-\alpha)\%$ confidence interval (CI) for θ is given by:

$$\hat{\theta}_{jack} \pm t_{\alpha/2,n-1}\,{}^{S}\!/_{\sqrt{n}}\,,$$

See Sahinler and Topuz [15] and Algamal and Rasheed [16].

5. Sensitivity analysis

Sensitivity represents deviation in model output with respect to changes in model's input(s). Global influence is the diagnostic influence depend on case deletion that represent one of the tools to perform sensitivity analysis introduced by Cook [11]. Case deletion is a popular method to investigate the influence of taking out the i^{th}case from the data on the parameter estimate. This method compare between $\hat{\theta}$ and $\hat{\theta}_{-i}$, where $\hat{\theta}_{-i}$ is MLE when the i^{th} case is deleted from original data. Then the i^{th} case could be considered as influential observation if $\hat{\theta}_{-i}$ is far from $\hat{\theta}$.

This methodology was conducted in many statistical models, see for example, Christensen, Pearson and Johnson [17], Davison and Tsai [18], Xie and Wei [19] and Xie and Wei [20]. The case deletion model for the LF regression model (4) is given by:

$$Y_J = \beta' x_i + \sigma Z_i;\;\; J = 1,2,3,\dots,n,\; J \neq i \qquad (7)$$

For model in (7), $\hat{\boldsymbol{\theta}}_{-i} = \left(\hat{\sigma}_{(i)}, \hat{\beta}_{(i)}\right)^T$ denote the ML estimator of $\boldsymbol{\theta}$ when i^{th} case is deleted. Then the generalized cook distance and likelihood distance are used to measure the effect of the i^{th} case on the ML estimator $\hat{\boldsymbol{\theta}} = \left(\hat{\sigma}, \hat{\beta}\right)^T$.

5.1. Generalized cook distance

Generalized cook distance is a method that measure global influence defined as the standardized norm of $\hat{\theta}_{-i} - \hat{\theta}$.

$$GD_i(\theta) = \left(\hat{\theta}_{-i} - \hat{\theta}\right)'\left\{\ddot{L}(\hat{\theta})\right\}\left(\hat{\theta}_{-i} - \hat{\theta}\right),$$

where $\ddot{L}(\hat{\theta})$ is the observed information matrix.

5.2. Likelihood distance

The likelihood distance is another method to measure the difference between $\hat{\theta}$ and$\hat{\theta}_{-i}$.

$$LD_i(\theta) = 2\left\{l(\hat{\theta}) - l(\hat{\theta}_{-i})\right\},$$

where $l(\hat{\theta})$ is a log likelihood function of θ from original data and $l(\hat{\theta}_{-i})$ is log likelihood function of θ when i^{th} case is deleted from original data.

6. Analysis of residual

Residuals analysis is an important method for checking the appropriateness of the proposed regression model. This will include studying any departure from error assumption and examine any existence of outliers. Several methods for residuals analysis were. introduced in the literature such as Collett [21]. In this study, we will concentrate on the Marginal and deviance methods.

6.1. Martingale residual

Martingale residual was proposed by Barlow and Prentice [22]. It is defined as the difference between the observed number of deaths and the expected in the interval $(0,t_i)$ and can be written as:

$$r_{M_i} = \delta_i + log\, S_Y(y_i, \hat{\theta}),$$

where δ_i is the censor indicator that takes 0 if censored and 1 if lifetime and $S_Y(y_i, \hat{\theta})$ is survival function for LF regression model. Then the martingale residual for LF regression model can be written as:

$$r_{M_i} = \begin{cases} 1 + log\{1 - exp[-exp(-\hat{z}_i)]\} & if\ i \in life\ time \\ log\{1 - exp[-exp(-\hat{z}_i)]\} & if\ i \in censored \end{cases},$$

where r_{M_i}takes the range between $-\infty$ and 1 but not symmetrically distributed (skewed). Therefore, transformation of the martingale residual will be used to reduce the skewness.

6.2. Deviance residual

Therneau, Grambsch and Fleming [23] introduced the deviance residuals for Cox model with no time-dependent explanatory variables. This residual is more symmetrically about zero from martingale residual and is given by:

$$r_{D_i} = sign(r_{M_i})\left\{-2\left[r_{M_i} + \delta_i\, log\left(\delta_i - r_{M_i}\right)\right]\right\}^{\frac{1}{2}}$$

In parametric regression models, the previous r_{D_i} is not a component of the deviance but can be used as a transformation of the martingale residual. Therefore, the deviance residual for LF regression model can be written as:

$$r_{D_i} = \begin{cases} sign(1 + \; log\{1 - exp[-exp(-\hat{z}_i)]\}) \\ \left\{-2\begin{bmatrix} 1 + log\{1 - exp[-exp(-\hat{z}_i)]\} + \\ log(-log\{1 - exp[-exp(-\hat{z}_i)]\}) \end{bmatrix}\right\}^{\frac{1}{2}} \\ if\ i \in life\ time \\ sign(\; log\{1 - exp[-exp(-\hat{z}_i)]\}) \\ \{-2[log\{1 - exp[-exp(-\hat{z}_i)]\}]\}^{\frac{1}{2}} \\ if\ i \in censored \end{cases}$$

7. Simulation study

Monte Carlo simulation study is conducted to estimate model parameters $\sigma\ and\ \beta_1$ when $\beta_0 = 2$ is fixed using ML and jackknife methods. This simulation was implemented for different sample sizes, n = 30, 50, and 100, from Fréchet distribution with the parameters $\lambda = 1.5$ and $\gamma = 10$ using various percentages of censoring 10, 30 and 50. Table 1 displays ML and jackknife parameters estimates along with the corresponding standard error (SE) for the LF regression model given in (4).

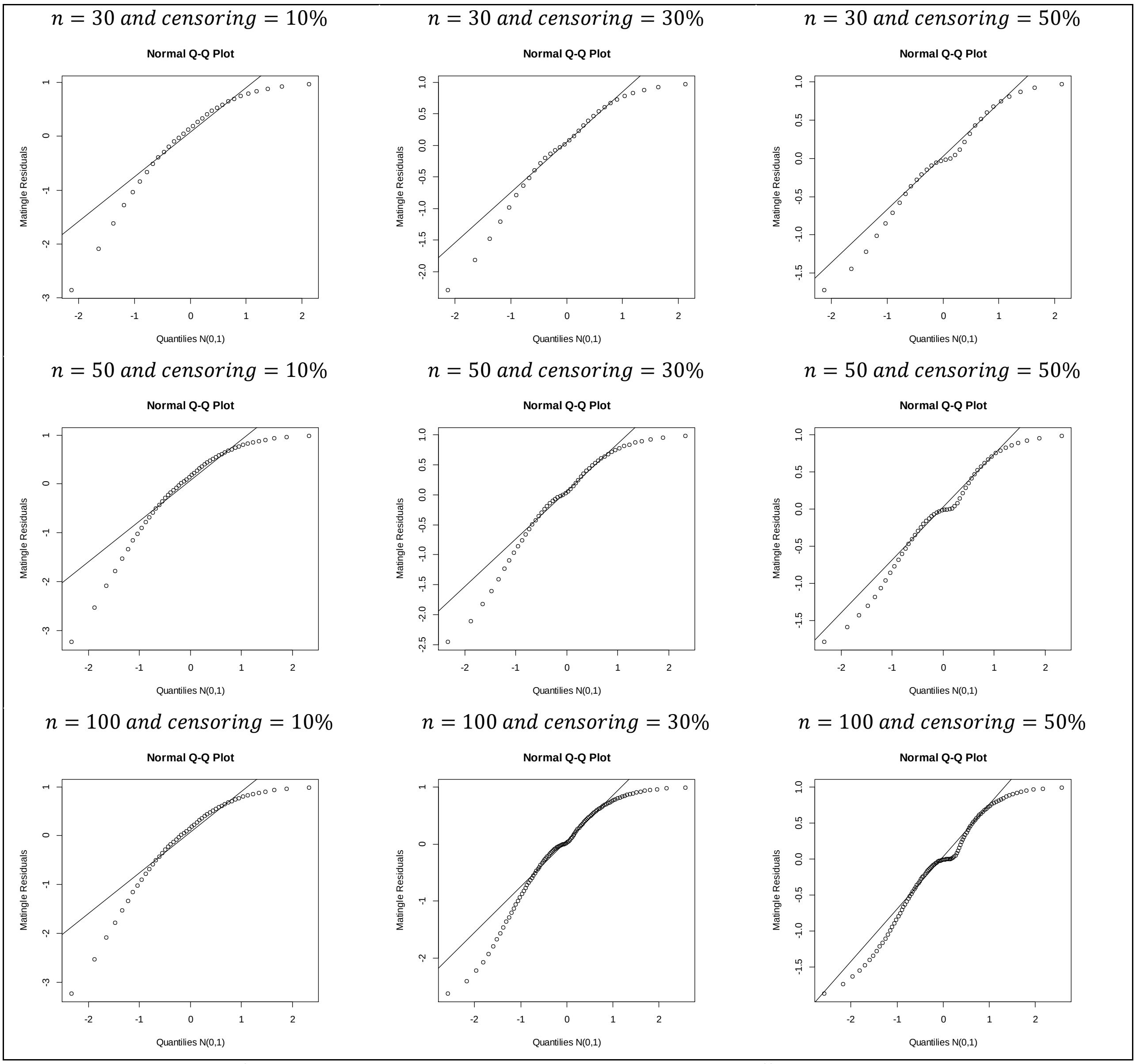
From these results, it can be noted that:

- SE of estimates using ML method is smaller compared to jackknife method.
- As percentage of censoring increased, the SE of parameter estimates increases at the same sample size.
- SE of parameter estimates decreases when the sample size increases.

In addition, simulation study is preformed to examine the form of the empirical distribution of $r_{M_i}$and r_{D_i} residuals at the different sample sizes and percentages of censoring. For 1000 samples generated, ML estimates are obtained for the parameters σ, β_1 when $\beta_0 = 2$ is fixed and r_{M_i} and r_{D_i} residuals are calculated. Then a plot of the mean ordered residuals versus the expected quantiles of the standard normal distribution is displayed in Fig. 2 and 3. Appendix B illustrates the algorithm that computes the ML estimates and residuals for LF regression model.

Table 1: Maximum Likelihood and Jackknife Estimates of the Parameters, SE for the LF Regression Model with $\sigma = 1/1.5$ and Fixed $\beta_0 = 2$.

			MLE		Jackknife	
n	%censoring	Parameter	Estimate	SE	Estimate	SE
30	10	σ	0.6462	0.1006	0.6629	0.1036
		β_1	0.4694	0.2149	0.4371	0.2399
	30	σ	0.6379	0.1111	0.6557	0.1153
		β_1	0.4663	0.2248	0.4311	0.2542
	50	σ	0.62090	0.1265	0.6414	0.1344
		β_1	0.4586	0.2418	0.4111	0.2793
50	10	σ	0.6533	0.0788	0.6626	0.0802
		β_1	0.4578	0.1680	0.4360	0.1823
	30	σ	0.6468	0.0873	0.6553	0.0888
		β_1	0.4552	0.1764	0.4307	0.1923
	50	σ	0.6320	0.0997	0.6445	0.1036
		β_1	0.4455	0.1902	0.4194	0.2102
100	10	σ	0.6593	0.0560	0.6648	0.0562
		β_1	0.4512	0.1193	0.4402	0.1264
	30	σ	0.6544	0.0622	0.6601	0.0624
		β_1	0.4475	0.1253	0.4380	0.1332
	50	σ	0.6442	0.0716	0.6525	0.0724
		β_1	0.4431	0.1358	0.4303	0.1453



$n = 30\ and\ censoring = 10\%$

$n = 30\ and\ censoring = 30\%$

$n = 30\ and\ censoring = 50\%$

$n = 50\ and\ censoring = 10\%$

$n = 50\ and\ censoring = 30\%$

$n = 50\ and\ censoring = 50\%$

$n = 100\ and\ censoring = 10\%$

$n = 100\ and\ censoring = 30\%$

$n = 100\ and\ censoring = 50\%$

Fig. 2: Normal Probability Plots for the Martingale Residual at Sample Size n = 30, 50 and 100, Different Censoring Percentages =10, 30 and 50, and Parameter Value $\sigma = 1/1.5$ and Fixed $\beta_0 = 2$.

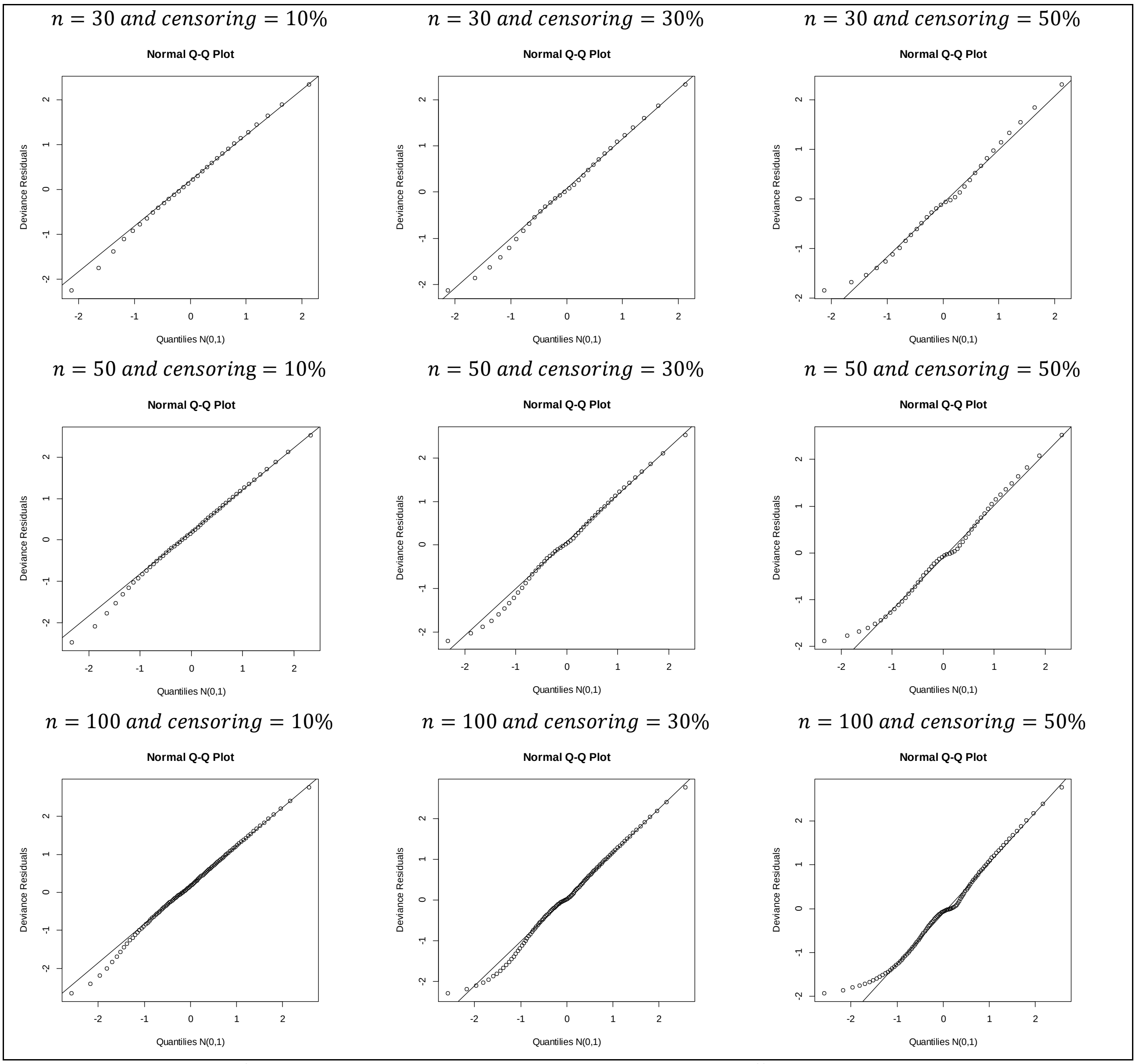


Fig. 3: Normal Probability Plots for the Deviance Residual at Sample Size *n* = 30, 50 and 100, Different Censoring Percentages = 10, 30 and 50, and Parameter Value $\sigma = 1/1.5$ and Fixed $\beta_0 = 2$.

From Fig. 2 and 3, we conclude the following:

- The empirical distribution of the deviance residual has close agreement to the standard normal distribution (SND) compared to the martingale residual.
- As the censoring percentage increases, the empirical distribution of the deviance residual moves away from the SND.
- As the sample size increases, the empirical distribution of the deviance residual approaches the SND.

8. Application

The Stanford heart transplant data is displayed in Kalbfleisch and Prentice [24] is used to illustrate the performance of the LF regression model. This data represents the survival time of 103 patients since acceptance into transplant program to death. The explanatory variables for each patient consist of age of patient at acceptance and two binary variables prior surgery and transplant.

The following variables in the study are:
t_i: Survival time (in days).
y_i: log survival time (in days).
$status_i$: Censoring indicator (0=censoring, 1=dead).
x_{i1}: Age of patients (in years).
x_{i2}: Previous surgery (0=No, 1=Yes).
x_{i3}: Transplant (0=No, 1=Yes).

8.1. Model validation

To check model validity, a plot of the empirical survival function by Kaplan Meier (KM) estimates and the estimated survival function based on fitting the Fréchet model is shown in Fig. 4. It can be seen from Fig. 4 that the logarithms of the times to event follow the LF distribution.
Therefore, the model fitted can be written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \sigma z_i,\ i = 1,2,\dots,103, \qquad (8)$$

where y_i follows the LF distributions given in (2).

8.2. Maximum likelihood and Jackknife estimation

Maximum likelihood and jackknife methods are used to estimate model parameters using *nlminb* function in *R* program. SE, 95% CI and p-value for each parameter are computed. The results are shown in the Table 2 and 3. It can be observed that the explanatory variables x_1 and x_3 are significant for the model at the significance level 5% for each method but x_2 is not significant. Also, the estimates from the two methods appear to be very similar.

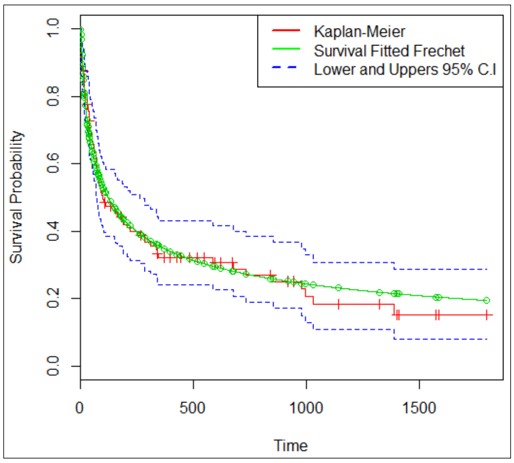


Fig. 4: Plot the Survival Function by Fitting the Fréchet Distribution and KM Function for Leukemia Data.

8.3. Global influence

The results of influence measure index plots using heart transplant data for LF regression models are shown in Fig. 5. It is clear that observations 15, 38 and 74 are possible influential observations in LF regression model.

8.4. Analysis of residual

The deviance residual for the fitted model is represented in Fig.6 (A). It indicates that all observations fall on the interval $(-3, 3)$ except observation 38 and are distributed randomly about zero. Therefore, it is expected that observation 38 is outliers.

8.5. Goodness of fitting

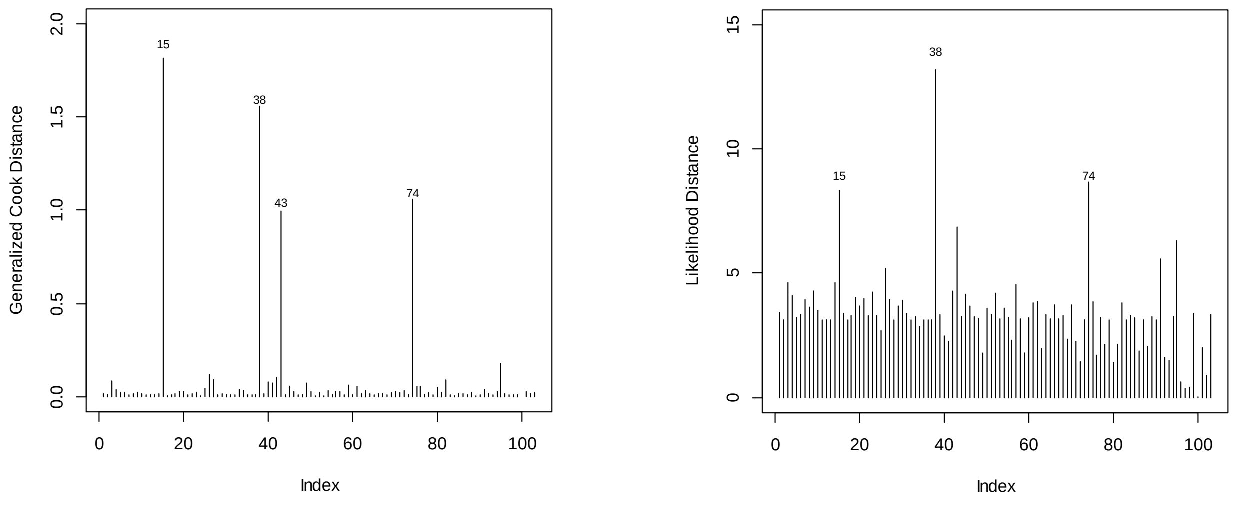
Fig.6 (B) represents the normal plot for deviance residual with a generated envelope simulation that illustrated in Appendix C. This plot shows that the LF model is suitable to fit the data, since all observations located inside the envelope.

Table 2: The ML Estimates for the Parameters of the LF Regression Model.

Parameter	Estimate	SE	95% CI	p-value
σ	1.7457	0.1484	(1.4548 , 2.0366)	-
β_0	4.2129	0.9153	(2.4189 , 6.0069)	< 0.001
β_1	$-$ 0.0431	0.0189	($-$0.0801 , $-$ 0.0061)	0.023
β_2	0.6902	0.5034	($-$ 0.2965 , 1.6769)	0.170
β_3	2.6572	0.3782	(1.9159 , 3.3985)	< 0.001

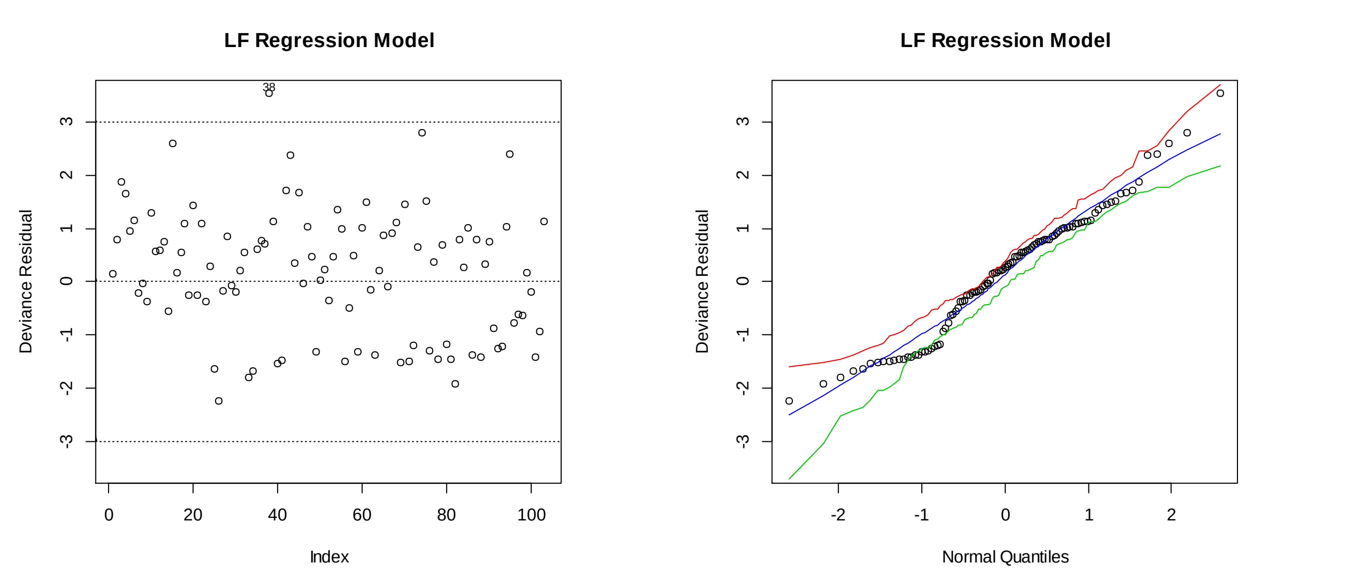
Table 3: The Jackknife Estimates for the Parameters of the LF Regression Model.

Parameter	Estimate	SE	95% CI
σ	1.8237	0.1839	(1.4633 , 2.1841)
β_0	4.2039	0.9893	(2.2643 , 6.1429)
β_1	$-$ 0.0436	0.0211	($-$ 0.0850 , $-$ 0.0022)
β_2	0.5695	0.6821	($-$ 0.7674 , 1.9064)
β_2	2.7046	0.4555	(1.8118 , 3.5974)



(A) (B)

Fig. 5: Plot Index of (A) Generalized Cook Distance and (B) Likelihood Distance for LF Regression Model.



(A) (B)

Fig. 6: (A) Index Plot of Deviance Residual. (B) Normal Probability Plot for the Deviance Residual with Envelopes.

8.6. Comparison between log-weibull and LF regression models

In Cruz, Ortega and Cordeiro [8], analysis has been conducted on the previous heart transplant survival time data under log-Weibull regression model. Therefore, we conduct a comparison between log-Weibull and LF regression model based on AIC and BIC criteria. Table 4 displays the results of these criteria which show that the LF regression model is more appropriate model compared to the log-Weibull regression model with smaller values of AIC and BIC.

Table 4: Statistics AIC and BIC for Comparing the Log-Weibull and LF Regression Models.

Model	AIC	BIC
Log-Weibull	353.4208	366.5944
LF	349.1578	362.3314

8.7. Final model

Based on this analysis, we conclude that the LF regression models are more appropriate for fitting these data compared to log-Weibull. Moreover, it can be noted that β_2 is not significant for this model at the level of 5%, see Table 2. Therefore, the final model fitted is:

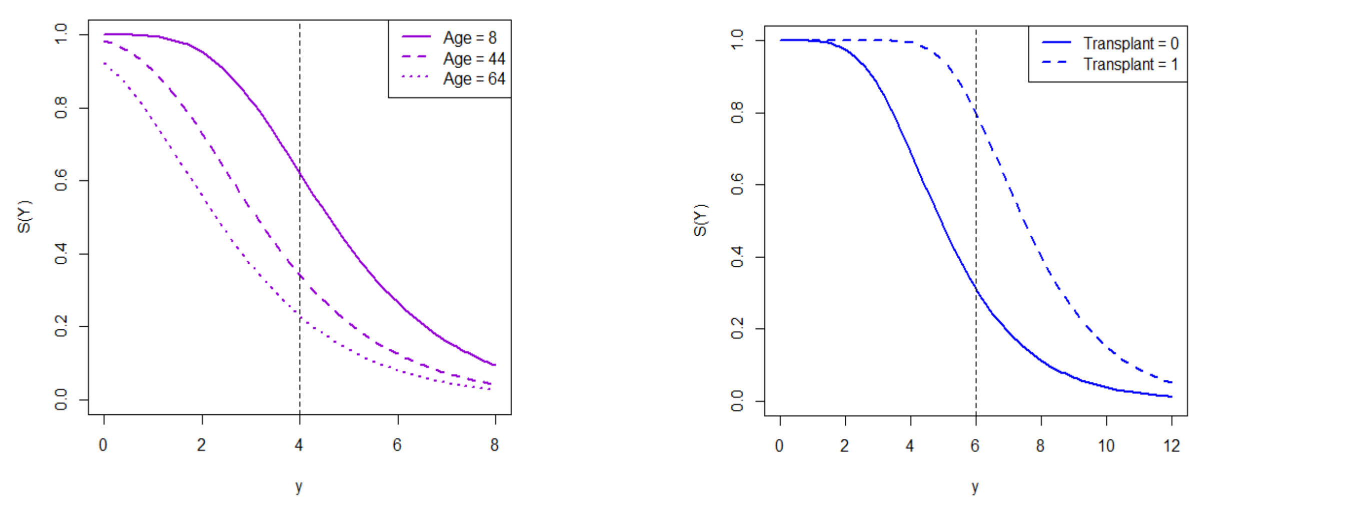
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + \sigma z_i, i = 1,2,\ldots,103, \qquad (9)$$

where $\mathrm{y_i}$ follows the LF given in (2).

Table 5 is represents the ML estimates of the parameters in the final model (9). It can be concluded that the log survival time increases for young patients who have received a heart transplant. Fig. 7 represents the survival function corresponding to explanatory variables for the fitted LF regression model.

Table 5: The ML Parameter Estimates of the LF Regression Model – Final Model.

Parameter	Estimate	SE	p-value	95% CI
σ	1.7492	0.1498	-	(1.4555 , 2.04288)
β_0	4.2728	0.9029	< 0.001	(2.50314 , 6.04239)
β_1	-0.0410	0.0186	0.028	$(-0.0774, -0.0045)$
β_3	2.5562	0.3696	< 0.001	(1.8319 , 3.2805)



(A) (B)

Fig. 7: Fitted Survival Functions from the LF Regression Model for the Heart Transplant. (A) for x_1 = Age, (B) for x_3 = Transplant.

From Fig. 7(A), it can be seen that $\hat{S}(4|\text{age} = 8) = 0.6205$, which means that approximately 62% of the patients at age equal to 8 years will be alive at y = 4 (≈ 55 days). Moreover, for patients at age to 44 and 64, $\hat{S}(4|\text{age} = 44) = 0.3408$ and $\hat{S}(4|\text{age} = 64) = 0.2295$, that is, the percentages of the patients that will be alive at y = 4 decreased to 34% and 23%, respectively. Similarly, Fig. 7(B), it can be noted that $\hat{S}(6|\text{transplant} = 0) = 0.3110$, which means that about 31% of the patients who did not receive a transplant will be alive at y = 6 (≈ 403 weeks), while for patients who received a transplant, $\hat{S}(6|\text{transplant} = 1) = 0.7994$, the survival percentage increases approximately to 80% at y = 6.

9. Concluding remarks

In this article, a LF regression model with right censored lifetime data is introduced. ML and jackknife methods were used to estimate model parameters. In addition, Monte Carlo simulation is carried out and has indicated that the empirical distribution of the deviance residual approaches the standard normal distribution. Moreover, the robustness features of the ML estimator from the fitted LF regression model are discussed through residuals and sensitivity analysis. The heart transplant data is used to illustrate the performance of the LF regression model. Goodness of fit is conducted for the data by constructing normal probability plot with simulated envelope where we observe that almost all observations fall within the envelope. Finally, the results of analysis showed that the proposed model provided more flexible and appropriate fit for the heart transplant data compared with the log-Weibull regression model using AIC and BIC criteria.

References

[1] J.F. Lawless,Statistical models and methods for lifetime data. 2th Ed, John Wiley and Sons, New Jersey. 2003.

[2] G.O. Silva, E. M. Ortega, V. G. Cancho, M. L. Barreto. Log-Burr XII regression models with censored data. Computational Statistics and Data Analysis 52(7) (2008) 3820-3842. https://doi.org/10.1016/j.csda.2008.01.003.

[3] J.M. Carrasco, E.M. Ortega,G.A. Paula, Log-modified Weibull regression models with censored data: Sensitivity and residual analysis. Computational Statistics and Data Analysis 52(8) (2008) 4021-4039. https://doi.org/10.1016/j.csda.2008.01.027.

[4] E.M. Ortega, G.M. Cordeiro, J.M. Carrasco, The log-generalized modified Weibull regression model. Brazilian Journal of Probability and Statistics0 (00) (2009) 1-29.

[5] E.M. Hashimoto, E.M. Ortega, V.G. Cancho, G.M. Cordeiro, The log-exponentiated Weibull regression model for interval-censored data. Computational Statistics and Data Analysis54 (4) (2009) 1017-1035. https://doi.org/10.1016/j.csda.2009.10.014.

[6] G.O. Silva, E.M. Ortega, V.G. Cancho, Log-Weibull extended regression model: estimation, sensitivity and residual analysis. Statistical Methodology7 (6) (2010) 614-631. https://doi.org/10.1016/j.stamet.2010.05.004.

[7] E.M. Hashimoto, E.M. Ortega, G.M. Cordeiro, M.L. Barreto, The Log-Burr XII Regression Model for Grouped Survival Data. Journal of biopharmaceutical statistics 22(1) (2012) 141-159. https://doi.org/10.1080/10543406.2010.509527.

[8] J.N.d.Cruz, E.M. Ortega, G.M. Cordeiro, The log-odd log-logistic Weibull regression model: modelling, estimation, influence diagnostics and residual analysis. Journal of Statistical Computation and Simulation 86(8) (2016) 1-23. https://doi.org/10.1080/00949655.2015.1071376.

[9] R.R. Pescim, E.M. Ortega, G.M. Cordeiro, M. Alizadeh, A new log-location regression model: estimation, influence diagnostics and residual analysis. Journal of Applied Statistics (2016) 1-20.

[10] S.Kotz,S. Nadarajah, Extreme value distributions: theory and applications, World Scientific,London, 2000. https://doi.org/10.1142/p191.

[11] R.D. Cook, Detection of influential observation in linear regression. Technometrics (1977) 15-18.

[12] R.G. Miller, An unbalanced jackknife. The Annals of statistics2 (5) (1974) 880-891. https://doi.org/10.1214/aos/1176342811.

[13] D.Tu, J. Shao, The Jackknife and bootstrap, Springer-Verlag, New York, 1995.

[14] H. Abdi, L. Williams, Jackknife. Encyclopedia of Research Design, Thousand Oaks, CA: Sage, (2010) 1-10.

[15] S. Sahinler,D. Topuz, Bootstrap and jackknife resampling algorithms for estimation of regression parameters. Journal of Applied Quantitative Methods2 (2) (2007) 188-199.

[16] Z.Y. Algamal, K.B. Rasheed, Re-sampling in Linear Regression Model Using Jackknife and Bootstrap. Iraqi Journal of Statistical Science18 (2010) 59-73.

[17] R. Christensen, L.M. Pearson, W. Johnson, Case-deletion diagnostics for mixed models. Technometrics 34(1) (1992) 38-45. https://doi.org/10.2307/1269550.

[18] A.Davison,C.-L. Tsai, Regression model diagnostics. International Statistical Review60 (3) (1992) 337-353. https://doi.org/10.2307/1403682.

[19] F.-C.Xie,B.-C. Wei, Diagnostics analysis for log-Birnbaum–Saunders regression models. Computational Statistics and Data Analysis51 (9) (2007a) 4692-4706. https://doi.org/10.1016/j.csda.2006.08.030.

[20] F.-C. Xie,B.-C. Wei, Diagnostics analysis in censored generalized Poisson regression model. Journal of Statistical Computation and Simulation 77(8) (2007b) 695-708. https://doi.org/10.1080/10629360600581316.

[21] D. Collett, Modelling survival data in medical research, CRC press, London, 2003

[22] W.E.Barlow,R.L. Prentice, Residuals for Relative Risk Regression, in Biometrika Biometrika Trust (1988) 65-74.

[23] T.M. Therneau, P.M. Grambsch, T.R. Fleming, Martingale-based residuals for survival models. Biometrika 77(1) (1990) 147-160. https://doi.org/10.1093/biomet/77.1.147.

[24] J.D.Kalbfleisch,R.L. Prentice, The statistical analysis of failure time data. 2thEd, John Wiley & Sons,New Jersey, 2002. https://doi.org/10.1002/9781118032985.

[25] Y.Zhao, A.H. Lee, K. K. Yau, G. J. McLachlan, Assessing the adequacy of Weibull survival models: a simulated envelope approach. Journal of Applied Statistics 38(10) (2011) 2089-2097. https://doi.org/10.1080/02664763.2010.545115.

Appendix A: Matrix of second derivatives $\ddot{\mathbf{L}}(\boldsymbol{\theta})$

$$\frac{\partial^2 l(\theta)}{\partial\sigma^2} = \frac{r}{\sigma^2} + \sum_{i=1}^{n} \tau_i\left[-2\,\ddot{z}_i - \dot{z}_i{}^2\, exp(-z_i) + 2\,\ddot{z}_i exp(-z_i)\right] + \sum_{i=1}^{n}(1-\tau_i)\left(\frac{\sigma(1-h_i)\left[-\dot{z}_iL_i + z_iL_i\left(\dot{z}_i - \dot{z}_i exp(-z_i)\right)\right] - z_iL_i[(1-h_i)+\sigma\,\dot{z}_iL_i]}{\sigma^2(1-h_i)^2}\right)$$

$$\frac{\partial^2 l(\theta)}{\partial\beta_k\partial\beta_j} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\tau_i\left[-\,x_{ij}\,x_{ik}\,exp(-z_i)\right] + \frac{1}{\sigma^2}\sum_{i=1}^{n}(1-\tau_i)\;x_{ij}\left\{\frac{(1-h_i)L_i\,x_{ik}[1-xp(-z_i)] - L_i{}^2\,x_{ik}}{(1-h_i)^2}\right\}$$

$$\frac{\partial^2 l(\theta)}{\partial\sigma\partial\beta_j} = \frac{-1}{\sigma^2}\sum_{i=1}^{n}\tau_i\,x_{ij}[1-\ exp(-z_i)] - \frac{1}{\sigma}\sum_{i=1}^{n}\tau_i\,x_{ij}\dot{z}_i\,exp(-z_i)\ - \frac{1}{\sigma^2}\sum_{i=1}^{n}(1-\tau_i)\left(\frac{x_{ij}L_i}{1-h_i}\right) + \frac{1}{\sigma}\sum_{i=1}^{n}(1-\tau_i)\,x_{ij}\left\{\frac{(1-h_i)L_i\dot{z}_i[1-\ exp(-z_i)] - L_i{}^2\dot{z}_i}{(1-h_i)^2}\right\},$$

where $z_i = (y_i - \beta^T x_i)/\sigma\,, h_i = exp[-exp(-z_i)]$,
$\dot{z}_i = (y_i - \beta^T x_i)/\sigma^2\,, \ddot{z}_i = (y_i - \beta^T x_i)/\sigma^3$,
$L_i = exp[-z_i - exp(-z_i)]$.

Appendix B: Algorithm ML estimation and residual analysis for the parameters of LF regression model

1) For a given values of the parameters $\lambda = 1.5$ and $\gamma = 10$, $T_1, T_2, \ldots, T_n$ are lifetimes generated from Fréchet distribution given in (1).
2) Generate the explanatory variable x_i from a standard uniform distribution.

3) Generate $C_1, C_2, \dots, C_n$ as a censoring time from uniform distribution $[0, \rho]$, where ρ was adjusted until obtaining the required censoring percentages.
4) The *log* lifetimes considered in each fit were calculated as $y = min\{log(T_i), log(C_i)\}$.
5) Estimate model parameter given in (4) by ML method using *nlminb* function in *R* program.
6) Compute the standard error for each estimate using *hessian* function in *R* program.
7) Compute the standardized, Martingale and deviance residuals.
8) The above steps are repeated 1000 times.
9) Plot the martingale and deviance residuals against the expected quantiles of normal distribution.

Appendix C: Algorithm of envelope simulation for normal probability plots for the deviance residual

1) Fit the LF regression model (4) to the observed lifetime data.
2) Using the parameter estimates obtained from the fitted model, will generate a sample of *n* independent observations.
3) Fit the model to generated sample in step (2) based on data set (τ_i, x_i).
4) Compute the values of deviance residuals and ordered them.
5) Repeat steps (2 – 4), *m* times.
6) Consider the *n* sets of the *m* ordered statistics, compute the mean, minimum, and maximum values across each set.
7) Plot these values and the ordered residuals of the original data against the normal scores.

The minimum and maximum values of the *m* ordered statistics constitute a simulated envelope to guide assessment of the model adequacy.
See, Ortega, Cordeiro and Carrasco [4]and Zhao et al. [25].