AJBAR Vol. 2(2), 2023: 30-39, ISSN: 2811-2881

Arid Zone Journal of Basic and Applied Research

Faculty of Science, Borno State University
Maiduguri, Nigeria

Journal homepage: https://www.azjournalbar.com

Research Article

A Study of the Mixture of Loglogistic-Loglogistic, Loglogistic-Gamma Distributions for the Analysis of Heterogenous Survival Data

Othman Musa Yakubu[1], Yusuf Abbakar Mohammed[1] and Akeyede Imam[2]
[1]Department of Mathematical sciences, University of Maiduguri
[2]Department of Mathematics, Federal University Lafia, Lafia
***Corresponding author**: othmanyakub@gmail.com, doi.org/10.55639/607. 171615

ARTICLE INFO:

Keyword:
Survival analysis,
Log-logistic,
Gamma,
MLE

ABSTRACT

Failure time data are used in survival analysis. The traditional parametric and nonparametric methods of survival analysis must be modified since the existence of censoring renders them inadequate. When one classical model may not be enough, parametric mixture models are used. To handle the heterogeneity of survival data, a more robust parametric mixture is required. For the study of survival data, this paper proposed a mixture of two distributions; the models are the Loglogistic-Loglogistic and Loglogistic-Gamma distributions. The models' performance was investigated using simulated data, and several iterations were run to test consistency. Expectation Maximization (EM) was employed to calculate the models' maximum likelihood parameters. The computed model parameters all fell within a narrow range of the postulated values. The models' consistency and stability were tested repeatedly through simulations using mean square error (MSE) and root mean square error (RMSE), and all were found to be stable and consistent. Real data were used to compare the fit of mixture models and classical distributions using information criteria (AIC). The best fit for the data was found using mixture models, which combine two different distributions. i.e., Loglogistic-Gamma distribution.

Corresponding author: Othman Musa Yakubu, Email: othmanyakub@gmail.com
Department of Mathematical sciences, University of Maiduguri

INTRODUCTION

Survival or reliability study is an area with its unique characteristic; it deals with the statistical methods of analysing survival data obtained from clinical studies of humans, laboratory study of animals and investigation of the durability of manufactured items, among other appropriate applications. Survival time can broadly be defined as the time to the occurrence of the event of interest, the event of interest can be the time to failure of a manufactured item, the time to occurrence of a disease, time to relapse, response to treatment, death, etc. The study of survival data has paid attention on predicting the probability of response, survival, or mean lifetime, comparing the survival distributions of experimental animals or, of human patients and the identification of risk and/or prognostic factors related to response, survival, and the development of a disease, (Lee and Wang, 2003). Parametric and nonparametric methods are usually employed (Kouassi and Singh, 1997), (Mohammed, *et al*, 2014).

Mixture models are being explored in survival and reliability analysis in recent times. Mixture models can be used to analyse failure-time data in a variety of ways. As a flexible way of modelling data, mixture models have several applications in situations where a single model may not suffice. They are applied where the data is heterogenous. Some authors proposed a mixture of three classical distributions (Mohammed *et al.*, 2014), to analyze survival data that has three different time overlapping phases (Blackstone *et al,* 1986).

Similarly, two component mixture models of different distributions were also studied (Blackstone *et al,* 1986), (Gupta, Akman, and Lvin 1999), therefore, this paper wish to enrich the study of two component, different distributions, mixture models for the analysis of survival data.

Survival analysis

Let *T* denote the survival time, which is a non-negative absolutely continuous random variable that represents the life time of individuals. If *F(t)* is the cumulative distribution of *T,* the survival function is defined to be;

$$S(t) = P(\text{an individual survives longer than } t) = p(T > t) = 1 - F(t) \quad (1)$$

The probability density function;

Similar to any other continuous random variable, the survival time *T* has a probability density function defined as the limit of the probability that an individual fails in the short interval *(t, t + Δt)* per unit width Δ*t.*

It can be expressed as;

$$f(t) = \lim_{\Delta \to t} \frac{P\{an\ individual\ dying\ in\ the\ interval(t,\ t+\Delta t)\}}{\Delta t} \quad (2)$$

The hazard function;

The hazard function *h* (t) of survival time T gives the conditional failure rate; it is defined as the probability of failure during small interval of time, given that the individual has survived to the beginning of the interval. It can be expressed as:

$$h(t) = \lim_{\Delta \to 0} \frac{P\{an\ individual\ of\ age\ t\ fails\ in\ the\ interval(t,\ t+\Delta t)\}}{\Delta t} \quad (3)$$

The hazard function *h* (t) can also be defined in terms of the cumulative distribution function *F* (t) and the probability density function $f(t)$

$$h(t) = \frac{f(t)}{1-F(t)} \quad (4)$$

The hazard function is also known as the *force of mortality, conditional mortality rate,* and *age-specific failure rate* (Lee and Wang 2003).

The survival function, *S*(t), probability density function, *f*(t), and hazard function, *h*(t), are mathematically equivalent.
The relationship is expressed thus:

$$h(t) = \frac{f(t)}{S(t)} \quad (5)$$

$$f(t) = \frac{d}{dt}[1 - S(t)] \quad (6)$$

$$h(t) = \frac{S'(t)}{S(t)} = -\frac{d}{dt} log_e\, S(t) \quad (7)$$

$$S(t) = \exp[-\int_0^t h(x)dx] \quad (8)$$

$$f(t) = h(t)\exp[-\int_0^t h(t)dx] \quad (9)$$

Gamma Distribution
The Gamma distribution has the pdf of the form:

$$f(t) = \frac{\lambda(\lambda t)^{k-1}e^{-\lambda t}}{\Gamma(k)} \quad t > 0 \quad (10)$$

Where k > 0 and t > 0 are parameters, λ^{-1} is the scale parameter and k is the shape parameter
The survival function 1 - *F*(t) is

$$S(t) = \int_t^{\infty} \frac{\lambda}{\Gamma(k)}(\lambda t)^{k-1}e^{-\lambda t}dx. \quad (11)$$

The hazard function,

$$h(t) \text{ is } f(t)/S(t) = \frac{\lambda(\lambda)^{n-1}}{(n-1)!\sum_{k=0}^{n-1}(1/k!)(\lambda t)^k} \quad (12)$$

it can be shown to be monotone increasing for k > 1.

Log-logistic Distribution
The log-logistic distribution is related to the logistic distribution in an identical fashion to how the log-normal and normal distributions are related with each other.
A logarithmic transformation on the logistic distribution generates the log-logistic distribution. Because of its flexible shapes, the log-logistic distribution has been illustrated to provide useful fits to data from many different fields, including engineering, economics, hydrology, and medical sciences.
The log logistic distribution is characterized by two parameters, γ (positive shape parameter) and η (positive scale parameter). A log-logistic random variable X with parameters γ and η has probability density function and survival function as follows;

$$f(x) = \frac{\Upsilon\eta x^{\eta-1}}{(1+\Upsilon x^{\eta})^2} \quad (13)$$

$$S(x) = \frac{1}{(1+\Upsilon x^{\eta})} \quad (14)$$

for $\Upsilon > 0$, $\eta > 0$. The log logistic distribution can be used to model the lifetime of an object, the lifetime of an organism, or a service time (Lee and Wang 2003).

The cumulative distribution function,

$$F(x) = P(X \leq x) = \frac{(\gamma x)^{\eta}}{1+(\gamma x)^{\eta}} \qquad x > 0 \qquad (15)$$

The hazard function is

$$h(x) = \frac{f(x)}{S(x)} = \frac{\gamma\eta(\gamma x)^{\eta-1}}{(1+(\gamma x))^{\eta}} \qquad x > 0 \qquad (16)$$

The cumulative hazard function,

$$H(x) = -\ln\big(S(x)\big) = \ln[1 + (\gamma x)^{\eta}] \qquad x > 0. \qquad (17)$$

The log-logistic distribution was used in survival analysis (Gupta, Akman, and Lvin 1999). Similarly, it was used to model economic data (Fisk, 1961).

METHODOLOGY

The Expectation maximisation algorithm (EM) was used to achieve the maximum likelihood estimation of the parameters of the model, it is a way to find maximum-likelihood estimates for model parameters when your data is incomplete, has missing data points, or has unobserved variables. It is an iterative way to approximate the maximum likelihood function. A model selection criterion based on the Akaike Information (AIC) was employed to find the mixture model that gives the best fit.

Data are simulated from a two-component parametric mixture of Loglogistic-Loglogistic, and Loglogistic-Gamma distributions. The model is evaluated by simulated data set, before it's applied to real dataset.

Parametric Mixture Model

Let $T_i, \ldots, T_n$ be n independent random variables, where T_j is the survival time of the j^{th} subject. We assume that the probability density function $f(t)$ of T_j is a mixture

$$f(t) = \sum_{i=1}^{a} \pi_i \, f_i(t;\theta) \qquad (18)$$

Where $f_i(t;\boldsymbol{\theta}_i)$ are the component densities of the mixture, $\boldsymbol{\theta}_i$ are the corresponding parameters of the ith density and the $\boldsymbol{\pi}_i$ are nonnegative probabilities that sum to one. That is,

$$\sum \pi_i = 1 \quad \text{and} \quad 0 \leq \boldsymbol{\pi}_i \leq 1 \qquad (i = 1,\ldots\ldots,a). \qquad (19)$$

The quantities $\boldsymbol{\pi}_1, \ldots_a$ are called mixing proportions or weights. Since the components $f_1(t;_1),\ldots,f_a(t;\boldsymbol{\theta}_a)$ are densities, the mixture (18) is a density.

It follows the survival function of failure-time T_j under mixture model is also a mixture,

$$S(t) = \sum \pi_i S_i(t;\theta_i) \qquad (20)$$

where $S_i(t;\theta)$ denotes the i^{ih} component survival function.

Loglogistic-Loglogistic Mixture Model

The mixture of log-logistic distribution can also be represented as the mixture of the component densities or sub groups as follows

$$f_{ll\text{-}ll}(t) = \pi f_{ll}(t;\gamma_1,\eta_1) + (1\text{-}\pi) f_{ll}(t;\gamma_2,\eta_2) \qquad (21)$$

where π is the mixing proportions and γ_1, γ_2, η_1, η_2, are the shapes and scales of the densities of the distribution.

- **Log-logistic-Gamma Mixture Model**

The mixture of the densities of Log-logistic and Gamma can be represented as

$$f_{ll\text{-}gm}(t) = \pi f_{ll}(t;\gamma_1,\eta_1) + (1\text{-}\pi) f_{gm}(t;\lambda_2,k_2) \qquad (22)$$

where is the weights of the mixture and γ_1, η_1 are the shape and scale of the loglogistic component and λ_2, k_2 are the shape and scale of the gamma component of the distribution.

The Expectation Maximization (EM) algorithm (McLachlan and Peel 2000), is a general approach to maximum likelihood estimation for problems of finite mixture models (Dempster, Laird, and Rubin 1977). Starting value initialisation is very important in the EM algorithm because the likelihood surface of mixture models tend to have

multiple modes (Zang, 2008). The EM algorithm typically produces improved result when started from reasonable initial values (Fraley and Raftery 2002).

The EM algorithm is a broadly applicable algorithm that provides an iterative procedure for computing Maximum Likelihood Estimates (MLE), (McLachlan and Peel 2000). Suppose the density of a random variable $\boldsymbol{Y}$ has an a component mixture form

$$f(y_i;\Psi)=\sum \pi_i f_i(y;\theta_i) \quad (23)$$

where $\Psi = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{a-1}, \boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_a)'$ is the vector containing all the unknown parameters in the mixture model. Let $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, _a)'$ be the vector of mixing proportions. Suppose $\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y_n}$ is an observed sample of size n, the likelihood for Ψ can be written as

$$L(\Psi) = \prod_{j=1}^{n} f(y_i;\Psi)$$
$$= \prod_{j=1}^{n} [\sum \pi_i f_i(y_i;\theta_i)_i] \quad (24)$$

Maximum likelihood estimation of mixture model is cumbersome to solve using the traditional method of taking derivative with respect to each parameter. These, and some other difficulties made modelling heterogeneous data unattractive for a long time. (McLachlan and Peel, 2000).

Mixture Model Selection

Regarding the mixture density estimation problem, the problem of determining the proper number of components and proper mathematical form of each component is faced. In other words, one need to determine which mixture model fits the data better. The question we try to answer here is a model selection problem.

The statistic Akaike Information Criterion (AIC) appears to be adequate for model selection in the mixture density estimation (McLachlan and Peel, 2000). Here, we follow the model selection approach using AIC proposed by Akaike, (Akaike, 1973).

$$\text{AIC} = -2logL(\Psi) + 2d \quad (25)$$

where d is the total number of independent parameters, n is the number of observation and $\boldsymbol{\Psi}$ is the estimate of the vector containing all the parameters.

RESULTS

The simulated data contains n = 100 observations. The Maximum Likelihood Estimates (MLE), of the parameters of each component of a mixture, and the Akaike Information Criteria (AIC) are used for model selection in each case, mean square error (MSE) and root mean square error (RMSE) are also employed, the MSE and RMSE are among the many ways to quantify the difference between an estimator and the true value of the quantity being estimated.

Loglogistic – Loglogistic

Data is simulated from a two-component parametric mixture of log-logistic distribution with n = 100 observations and the mixing proportions are λ_1=0.5, λ_2 =0.5 respectively.

Table 1 list the MLEs of the parameters of the mixture model, it can be seen that, the model is suitable for the estimation of the parameters because the estimated parameters are closed to the postulated values. Table 2 shows that, the model is also consistent and stable in its estimation.

Table 1: MLEs of Log-Logistic – Log-Logistic model with no repetitions

Log-logistic – Log-logistic						
Parameter	λ_1	λ_2	α_1	α_2	β_1	β_2
Postulate	0.5	0.5	10	10	0.7	0.3
Estimates	0.49	0.51	10.12	10.45	0.71	0.29

It can be seen from the table 1 above that, the model is suitable for the estimation of the parameters and table 2 shows that, it is equally consistent.

Table 2: MLEs of Log-Logistic – Log-Logistic model with 300 repetitions

Log-logistic – Log-logistic						
Parameter	λ_1	λ_2	α_1	α_2	β_1	β_2
Postulate	0.5	0.5	10	10	0.7	0.3
Estimates	0.52	0.48	10.13	10.36	0.70	0.30
MSE	7.22e-07	7.22e-07	4.17e-03	4.01e-03	1.14e-06	1.73 e-07
RMSE	0.0008	0.0008	0.0667	0.0638	0.0009	0.0004

Figure 1 compares the density function of each component and the density of the mixture model, it can be seen that, the mixture model fits the data better.

Simulations were repeated 300 times to assess the consistency of the model using the MSE and the RMSE obtained.

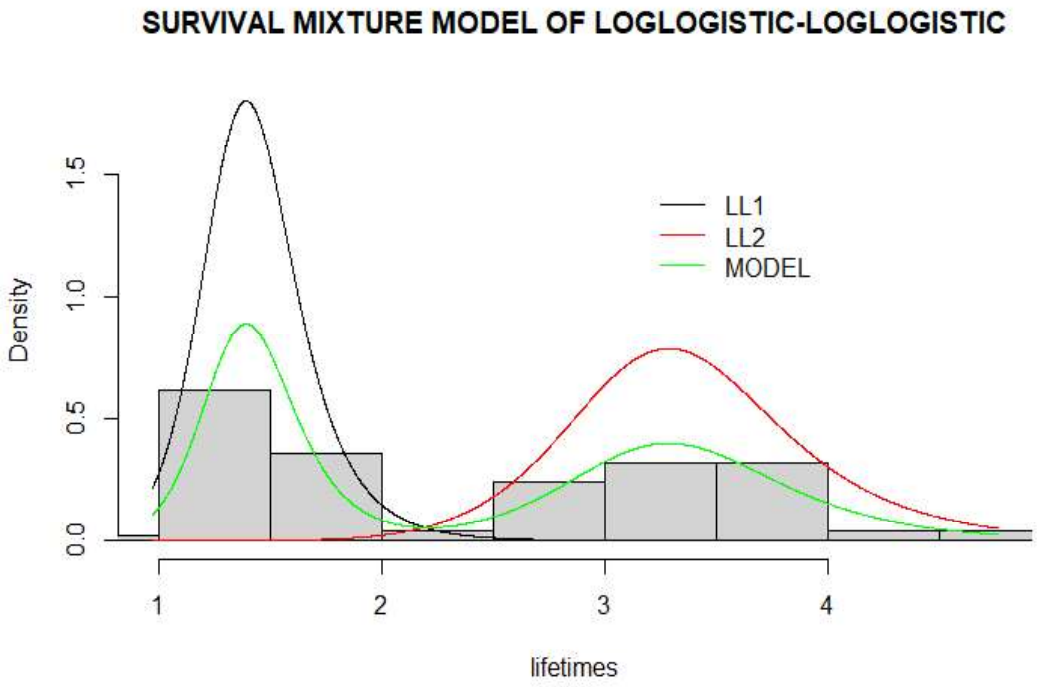

Figure 1 Density of survival mixture model of log-logistic – log-logistic model.

Loglogistic-Gamma

Table 3: MLEs of Log-Logistic-Gamma model with no repetitions.

Log-logistic-Gamma						
Parameter	λ_1	λ_2	α_1	α_2	β_1	β_2
Postulate	0.5	0.5	10	2	4	.1
Estimate	0.52	0.48	10.52	1.88	3.95	0.11

Table 3 shows that, the model estimates are very close to the postulated values, hence the model can be said to be efficient.

Table 4: MLEs of Log-Logistic - Gamma model with 300 repetitions

Loglogistic-Gamma						
Parameter	λ_1	λ_2	α_1	α_2	β_1	β_2
Postulate	0.5	0.5	10	2	4	0.1
Estimate	0.44	0.56	10.69	1.55	4.34	0.11
MSE	4.19e-06	4.19e-06	1.34e-02	2.24e-04	2.26e-03	1.43e-06
RMSE	0.002	0.002	0.116	0.015	0.0478	0.001

Simulations were repeated 300 times to test for consistency of the model and it is found to be consistent as can be seen from the MSE and RMSE in table 4, which are very close to zero.

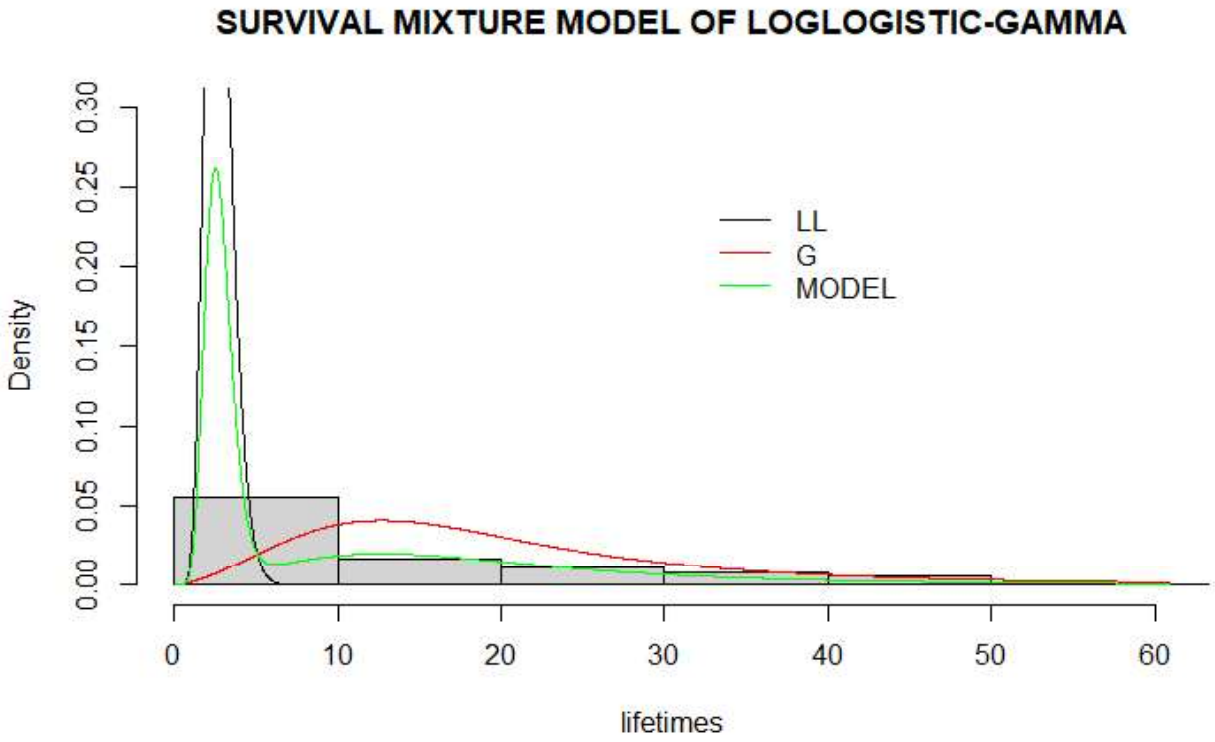

Figure 2: Density of survival mixture model of Gamma – log-logistic model.

The densities of the two classical distributions were also compared to the model as shown in figure 2 above, it can be seen that, the model fits the data better.

Real data application

The data used is the Acute Myelogenous Leukemia (aml) dataset in R statistical software. The question at the time was whether the standard course of chemotherapy should be extended ("maintained") by additional cycles.

Table 5: MLEs and AIC of aml dataset using the Loglogistic-Loglogistic model.

Model	Estimate		LL	AIC
Log-Logistic-Log-Logistic	$\hat{\lambda}_1 = 0.13$	$\hat{\lambda}_2 = 0.87$		
	$\hat{\alpha}_1 = 179.51$	$\hat{\alpha}_2 = 1.4$	-98.93	209.86
	$\hat{\beta}_1 = 0.17$	$\hat{\beta}_2 = 20.69$		

Table 5 shows the MLEs and the AIC of the Loglogistic-Loglogistic model applied on the real dataset, the aml dataset.

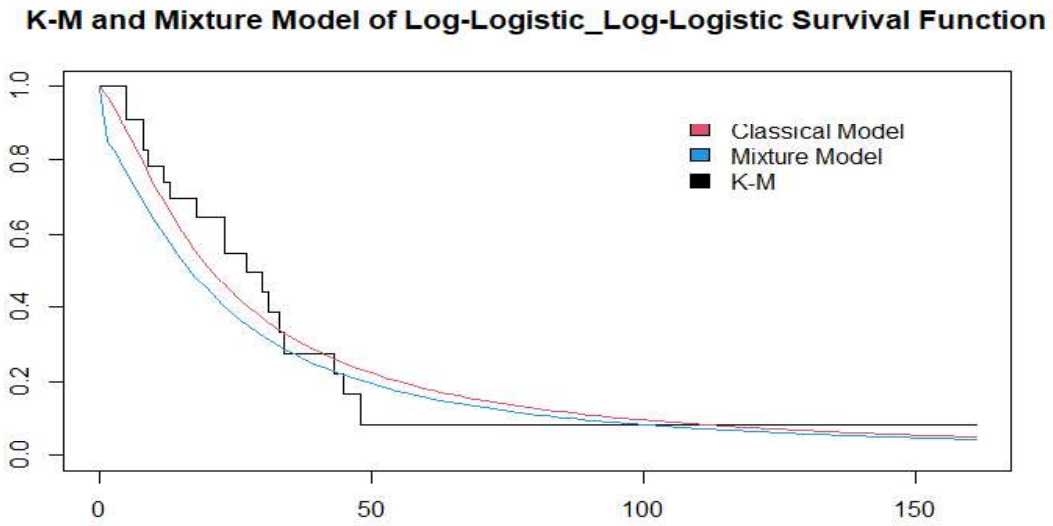

Figure 3: Kaplan Meier **(**K-M) Survival curve of Loglogistic-Loglogistic model.

Table 6: MLEs and AIC of Loglogistic-Gamma model.

Model	Estimate		LL	AIC
Loglogistic-Gamma	$\hat{\lambda}_1 = 0.13$	$\hat{\lambda}_2 = 0.87$		
	$\hat{\alpha}_1 = 179.51$	$\hat{\alpha}_2 = 1.4$	-98.84	209.69
	$\hat{\beta}_1 = 0.17$	$\hat{\beta}_2 = 20.69$		

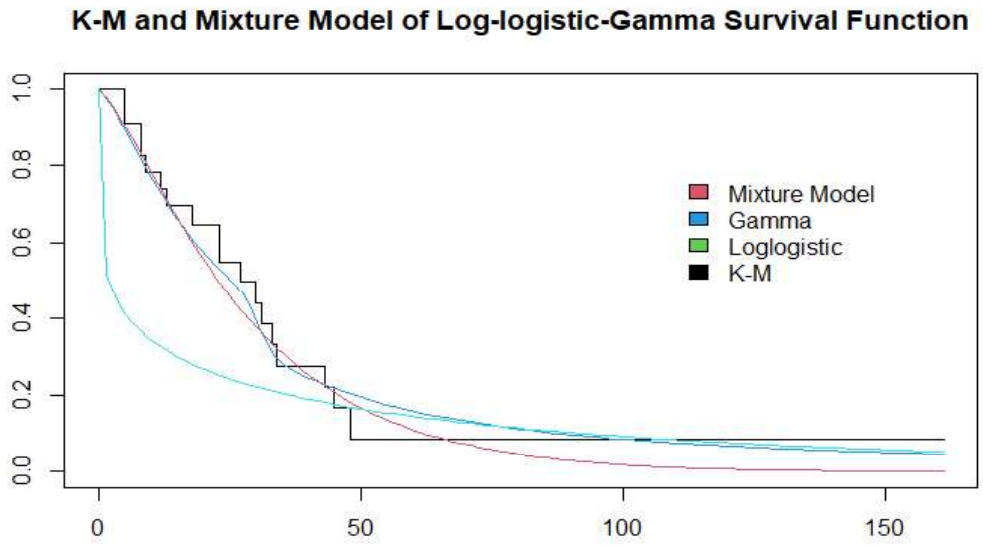

Figure 4: K-M Survival function of Loglogistic-Gamma model.

It is observed that the Loglogistic-Loglogistic model fits both the simulated and real data well, however, the Loglogistic-Gamma model gives the best fit to both datasets, because it has the lowest Akaike Information Criteria (AIC), and comparing the Kaplan Meier (K-M) survival curves it can also be seen that the K-M survival curve of the loglogistic-Gamma gives a better fit.

The estimation of the parameters of the model was successful in both the simulated and the real data, as the estimated values were relatively close to postulated values.

As can be seen, the mixture model of Loglogistic-Gamma gives the best fit for the real data among the proposed models.

Table 7: Compares the MLEs and AICs of the two models i.e., Loglogistic-Loglogistic and Loglogistic-Gamma models on real dataset (aml)

Model	Estimate		LL	AIC
Log-Logistic-Log-Logistic	$\hat{\lambda}_1 = 0.13$	$\hat{\lambda}_2 = 0.87$		
	$\hat{\alpha}_1 = 179.51$	$\hat{\alpha}_2 = 1.4$	-98.93	209.86
	$\hat{\beta}_1 = 0.17$	$\hat{\beta}_2 = 20.69$		
Loglogistic-Gamma	$\hat{\lambda}_1 = 0.13$	$\hat{\lambda}_2 = 0.87$		
	$\hat{\alpha}_1 = 179.51$	$\hat{\alpha}_2 = 1.4$	-98.84	209.69
	$\hat{\beta}_1 = 0.17$	$\hat{\beta}_2 = 20.69$		

CONCLUSION

The paper proposed a two-component mixture model of classical distributions, namely, Loglogistic-Loglogistic and Loglogistic-Gamma distributions to analyze heterogenous survival data, simulated and real data were employed to assess the performance of the models, and the models were found to estimate the parameters successfully as the estimates were close to the postulated values.

It is found that the mixture of two different distributions i.e., Loglogistic-Gamma gives the best fit to the real data applied.

REFERENCES

Akaike, H. (1973). Information theory and extension of the maximum likelihood principle. 2nd International Symposium on Information Theory, B.N. Petov and F. Csaki (eds), Akademiai Kiado, Budapest, 267-281.

Blackstone, E. H., Naftel, D. C., & Turner Jr, M. E. (1986). The decomposition of time-varying hazard into phases, each incorporating a separate stream of concomitant information. *Journal of the American Statistical Association*, *81*(395), 615-624.

Erisoglu, U. Erisoglu, M. Erol, H. (2011). Mixture model approach to the analysis of heterogeneous survival time data. *Pakistan Journal of Statistics*, *28*(1), 115-130.

Erişoğlu, Ü. Erişoğlu, M., & Erol, H. (2011). A mixture model of two different distributions approach to the analysis of heterogeneous survival data. *International Journal of Computational and Mathematical Sciences*, *5*(2), 75-79.

Fisk, P. R. (1961). The graduation of income distributions. *Econometrica: Journal of the Econometric Society*, 171-185.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, *97*(458), 611-631.

Gupta, R. C., Akman, O., & Lvin, S. (1999). A study of log-logistic model in survival analysis. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *41*(4), 431-443.

Kouassi, D. A., & Singh, J. (1997). A semiparametric approach to hazard estimation with randomly censored observations. *Journal of the American Statistical Association*, *92*(440), 1351-1355.

Lee, E. T., & Wang, J. (2003). *Statistical methods for survival data analysis* (Vol. 476). John Wiley & Sons.

McLachlan, G. J., & Krishnan, T. (2007). *The EM algorithm and extensions*. John Wiley & Sons.

Mohammed, Y. A., Yatim, B., & Ismail, S. (2013). A simulation study of a parametric mixture model of three different distributions to analyze heterogeneous survival data. *Modern Applied Science*, *7*(7), 1-9.

Mohammed, Y. A., Yatim, B., & Ismail, S. (2014). A parametric mixture model of three different distributions: An approach to analyse heterogeneous survival data. In *AIP Conference Proceedings*, *1605*(1), 1040-1045.

Peel, D. A. V. I. D., & MacLahlan, G. (2000). Finite mixture models. *John & Sons*.

Zhang, Y. (2008). *Parametric mixture models in survival analysis with applications*. Temple University.