Formulation of Kumaraswamy Generalized Inverse Lomax Distribution

A B N Manurung[1,*], **S Nurrohmah**[2], **I Fithriani**[3]

[1] Mathematics Department, Universitas Indonesia, Depok, Jawa Barat, Indonesia

* Corresponding author's email: andrewbony.sj@gmail.com

Abstract. Lifetime data is a type of data that consists of a waiting time until an event occurs and modelled by numerous distributions. One of its characteristics that is interesting to be studied is the hazard function due to the flexibility that it has compared to other characteristics of distribution. Inverse Lomax (IL) distribution is one of the distributions considered to have advantages in modelling hazard shape and extended in several ways to address the problem of non-monotone hazard which is often encountered in real life data. However, it needs to be extended to another family of distribution to increase its modelling potential and Kumaraswamy Generalized (KG) family of distribution is used as it adds two more parameters to the distribution. The newly developed distribution is called the Kumaraswamy Generalized Inverse Lomax (KGIL) distribution. The main characteristics of KGIL distribution will be derived, such as cumulative distribution function (cdf), probability density function (pdf), hazard function, and survival function. Maximum likelihood method will also be used to estimate the parameters. The application of the new model is based on head-and-neck cancer lifetime data set. The modelling results show that the KGIL distribution is the best to capture important details of the data set considered.

1. Introduction

Probability distribution has helped to model so many uncertainty problems on our daily basis. The most well-known probability distribution is normal distribution. Normal distribution could be used to model the distribution of human IQ scores, body height, and many others [1]. Normal distribution is a good distribution to depict a symmetry probability distribution. However, probability distribution encountered in daily life is not always symmetric. Asymmetric distribution is called skewed distribution since it could have skewed parts in its pdf curve. Skewness coefficient is used to describe how skewed a distribution is [2]. When dealing with this type of distribution, symmetric probability distributions cannot be used anymore. Some new distributions are developed to give a better picture for an asymmetric distributed data.

Kumaraswamy distribution is one of the asymmetric probability distributions. It was developed to model hydrology phenomena at first, such as daily precipitation and reservoir volume [3]. The cdf and pdf of the distribution are given respectively as (1) and (2) as

$$F(x;p,q) = 1-(1-x^p)^q, x \in [0,1], p>0, q>0 \quad (1)$$

and

$$f(x;p,q) = pqx^{p-1}(1-x^p)^{q-1}, x \in [0,1]; p,q>0 \quad (2)$$

with $p > 0$ and $q > 0$ are shape parameters.

It can be seen from the equations (1) and (2) that Kumaraswamy distribution has a support $x \in [0,1]$ and it suits the range of probability values. Therefore, Cordeiro et al. [4] proposed a new family of distributions using Kumaraswamy distribution to extend another distribution into having a more flexible characteristic in its modelling potential and hazard shapes. The newly proposed family of distribution is called Kumaraswamy Generalized (KG) family distribution. Its cdf and pdf are given in (3) and (4) as

$$F(x) = 1 - (1 - G(x)^p)^q, x > 0; p, q > 0 \tag{3}$$

and

$$f(x) = pq(1 - G(x)^p)^{q-1} G(x)^{p-1} g(x), x > 0; p, q > 0 \tag{4}$$

with $G(x)$ and $g(x)$ are respectively pdf and cdf of the baseline distribution.

Inverse Lomax (IL) distribution is one of the newly developed distributions with a good feature to describe asymmetric data. It is also known to be extended in several ways to capture non-monotone problems encountered in real life data [5]. The pdf and cdf of IL distribution are respectively given in (5) and (6) below:

$$G(x; \lambda, \beta) = \left(1 + \frac{\beta}{x}\right)^{-\lambda}, x > 0;\ \lambda, \beta > 0 \tag{5}$$

and

$$g(x; \lambda, \beta) = \lambda \beta x^{-2} \left(1 + \frac{\beta}{x}\right)^{-(\lambda+1)}, x > 0;\ \lambda, \beta > 0 \tag{6}$$

where $\lambda > 0$ is shape parameter and $\beta > 0$ is scale parameter.

2. Kumaraswamy Generalized Inverse Lomax (KGIL) Distribution

A random variable X follows KGIL distribution if its pdf and cdf could be expressed respectively in equation (7) and (8) as follows:

$$F(x; \beta, \lambda, p, q) = 1 - \left[1 - \left(1 + \frac{\beta}{x}\right)^{-\lambda p}\right]^q, x > 0;\ \lambda, \beta, p, q > 0 \tag{7}$$

and

$$f(x; \beta, \lambda, p, q) = \lambda \beta p q x^{-2} \left(1 + \frac{\beta}{x}\right)^{-[\lambda p+1]} \left[1 - \left(1 + \frac{\beta}{x}\right)^{-\lambda p}\right]^{q-1}, \quad x > 0;\ \lambda, \beta, p, q > 0 \tag{8}$$

where β is scale parameter and p, q, λ are scale parameters [6].

From the pdf and cdf above, we can derive some of the main characteristics of KGIL distribution:

2.1. Survival function

Survival function is the complementary of cdf. Therefore, from the KGIL cdf (7), survival function could be obtained as:

$$\begin{aligned} S(x; \beta, \lambda, p, q) &= 1 - F(x; \beta, \lambda, p, q) \\ &= \left[1 - \left(1 + \frac{\beta}{x}\right)^{-\lambda p}\right]^q, x > 0;\ \lambda, \beta, p, q > 0 \end{aligned} \tag{9}$$

2.2. Hazard function

Hazard function of KGIL distribution could be derived from the mathematical definition of hazard function:

$$h(x) = \frac{f(x)}{S(x)}$$
$$= \frac{\lambda\beta pq x^{-2}\left(1+\frac{\beta}{x}\right)^{-[\lambda p+1]}\left[1-\left(1+\frac{\beta}{x}\right)^{-\lambda p}\right]^{q-1}}{\left[1-\left(1+\frac{\beta}{x}\right)^{-\lambda p}\right]^{q}}; x>0; p,q,\lambda,\beta>0$$
$$= \frac{\lambda\beta pq x^{-2}\left(1+\frac{\beta}{x}\right)^{-[\lambda p+1]}}{1-\left(1+\frac{\beta}{x}\right)^{-\lambda p}}; x>0; p,q,\lambda,\beta>0 \quad (10)$$

3. Parameter Estimation of the KGIL Distribution

The MLE method is an approach used in determining the parameters that maximize the likelihood function of the sample data. Taking an observed sample $x_1, x_2, \ldots, x_n$ from the KGIL distribution, the corresponding likelihood function can be represented as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta})$$
$$= \prod_{i=1}^{n} \lambda\beta pq {x_i}^{-2}\left(1+\frac{\beta}{x_i}\right)^{-[\lambda p+1]}\left[1-\left(1+\frac{\beta}{x_i}\right)^{-\lambda p}\right]^{q-1}$$
$$= (\lambda\beta pq)^n \times \prod_{i=1}^{n} {x_i}^{-2} \times \prod_{i=1}^{n}\left(1+\frac{\beta}{x_i}\right)^{-[\lambda p+1]} \times \prod_{i=1}^{n}\left[1-\left(1+\frac{\beta}{x_i}\right)^{-\lambda p}\right]^{q-1} \quad (11)$$

The log likelihood function is given by:

$$l(\boldsymbol{\theta}) = \ln[L(\boldsymbol{\theta})]$$
$$= n(\ln\lambda + \ln\beta + \ln p + \ln q) - 2\sum_{i=1}^{n}\ln x_i - (\lambda p+1)\sum_{i=1}^{n}\ln\left(1+\frac{\beta}{x_i}\right) + (q - 1)\sum_{i=1}^{n}\ln 1-\left(1+\frac{\beta}{x_i}\right)^{-\lambda p} \quad (12)$$

Thus, the MLEs of p, q, λ, and β, respectively denoted by $\hat{p}, \hat{q}, \hat{\lambda}$, and $\hat{\beta}$, could be obtained by deriving its log likelihood partially to each parameter equal zero.

$$\frac{\partial}{\partial p}l(\boldsymbol{\theta}) = \frac{n}{p} - \lambda\sum_{i=1}^{n}\ln\left(1+\frac{\beta}{x_i}\right) + (q-1)\lambda\sum_{i=1}^{n}\frac{\left(1+\frac{\beta}{x_i}\right)^{-\lambda p}\ln\left(1+\frac{\beta}{x_i}\right)}{1-\left(1+\frac{\beta}{x_i}\right)^{-\lambda p}} \quad (13)$$

$$\frac{\partial}{\partial q}l(\boldsymbol{\theta}) = \frac{n}{q} + \sum_{i=1}^{n}\ln\left[1-\left(1+\frac{\beta}{x_i}\right)^{-\lambda p}\right] \quad (14)$$

$$\frac{\partial}{\partial\lambda}l(\boldsymbol{\theta}) = \frac{n}{\lambda} - p\sum_{i=1}^{n}\ln\left(1+\frac{\beta}{x_i}\right) + p(q-1)\sum_{i=1}^{n}\left[\frac{\left(1+\frac{\beta}{x_i}\right)^{-\lambda p}\ln\left(1+\frac{\beta}{x_i}\right)}{1-\left(1+\frac{\beta}{x_i}\right)^{-\lambda p}}\right] \quad (15)$$

$$\frac{\partial}{\partial\beta}l(\boldsymbol{\theta}) = \frac{n}{\beta} - (\lambda p+1)\sum_{i=1}^{n}\left[\frac{1}{x_i\left(1+\frac{\beta}{x_i}\right)}\right] + (q-1)\sum_{i=1}^{n}\left[\frac{\lambda p\left(1+\frac{\beta}{x_i}\right)^{-\lambda p}}{x_i\left(1+\frac{\beta}{x_i}\right)}\right] \quad (16)$$

It could be observed from equation (14) that the MLE of q satisfy the following simple equation:

$$\frac{\partial}{\partial q} l(\boldsymbol{\theta}) = \frac{n}{q} - \sum_{i=1}^{n} \ln\left[1 - \left(1 + \frac{\beta}{x_i}\right)^{-\lambda p}\right] = 0$$

$$\frac{n}{q} = \sum_{i=1}^{n} \ln\left[1 - \left(1 + \frac{\beta}{x_i}\right)^{-\lambda p}\right] \tag{17}$$

$$\hat{q} = -\frac{n}{\sum_{i=1}^{n} \ln\left[1 - \left(1 + \frac{\beta}{x_i}\right)^{-\lambda p}\right]} \tag{18}$$

4. Data Illustration

We will apply the KGIL distribution to a real-life data set from the study of head-and-neck cancer patients conducted by Northern Carolina Oncology Group [7]. In this chapter, we will also use Kolmogorov-Smirnov test to statistically prove that KGIL distribution is better than IL distribution in modelling the data set considered. The observation contains of 42 data on patient survival time until death occurs. Exploratory data analysis for the data set considered is presented in Table 1, including the mean, median, standard deviation, standard error, range, skewness, and kurtosis. The graph of total test time (TTT) plotted for the data presented in Figure 1.

Table 1. Descriptive statistics of head-and-neck cancer patients survival time.

Mean	Median	Standard deviation	Skewness	Kurtosis	Standard error	Range
280.167	160	303.125	2.409	5.921	46.773	1410

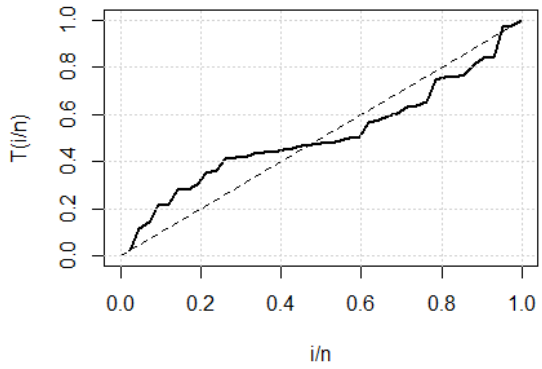

Figure 1. The graph of total test time (TTT) plots.

From the results in Table 1, we can see that the data set is positively skewed, and light-tailed with excess kurtosis of 5.921. In addition, we can see that the curve of TTT plots in Figure 1 is strictly concave then strictly convex. The mentioned shape of TTT curve tells us that the data set has an inverted-bathtub hazard shape, and the hazard shape will be shown in Figure 6.

4.1. Best model

In this subchapter, it will be shown that KGIL distribution fits the data set of head-and-neck cancer patients survival time better than IL distribution.

4.1.1. Comparison between both cdf of models and empirical cdf. From Figure, we can see that the blue line is the empirical CDF of the data set and it is best-fitted by the KGIL distribution. However, it is not enough to say that KGIL distribution can fit the data set better than IL distribution only by considering the graph of comparison between three cdfs. Hence, we have to do an objective analysis by comparing the values of AIC produced by IL and KGIL model.

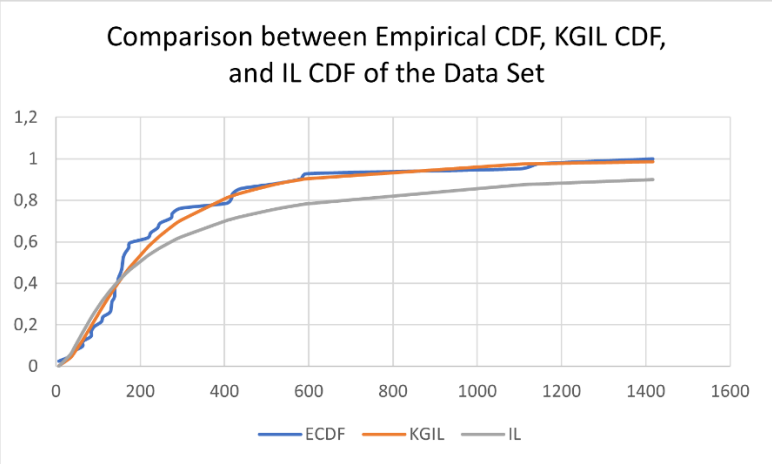

Figure 2. Comparison of three curves of CDF based on the data set.

4.1.2. Comparison by AIC values. In general, it is considered that the smaller the values of AIC, the better the fit of the model. Based on the AIC values, it can be concluded that the KGIL distribution fits the considered data set better than the IL distribution.

Table 2. AIC values of both models.

Distribution	AIC
Inverse Lomax	565,4261
KGIL	559,2302

4.1.3. Comparison by BIC values. The same thing applies to BIC values whereas the model with smaller BIC vales is considered a better model compared to the other one. Based on the AIC values, it can be concluded that the KGIL distribution fits the considered data set better than the IL distribution.

Table 3. BIC values of both models.

Distribution	BIC
Inverse Lomax	564.6725
-KGIL	557.7231

4.1.4. Likelihood Ratio Test. To provide more information about the best distribution to modelling the considered data set, we can perform likelihood ratio test (LRT) through these steps:

1. Formulating hypotheses
   H_o: Inverse Lomax distribution adequately describes the data
   H_1: Kumaraswamy Generalized Inverse Lomax distribution adequately describes the data
2. Calculating Likelihood Ratio Test (LRT) statistics
$$\begin{aligned} LRT &= -2 * (log - likelihood(KGIL) - log - likelihood(IL) \\ &= 561.4261 - 551.2302 \\ &= 10.1959 \end{aligned}$$
3. Degrees of freedom
$$\begin{aligned} df &= number\ of\ parameter\ of\ KGIL - number\ of\ parameter\ of\ IL \\ &= 4 - 2 \\ &= 2 \end{aligned}$$
4. Calculating the p-value with help of RStudio
   The corresponding p-value with $LRT = 10.1959$ and $df = 2$ is $p - value = 0.006109258$
5. Significance level $\alpha = 0.05$

6. Conclusion
   From this Likelihood Ratio Test, the p-value is less than the significance level. We reject the null hypothesis in favor of the IL distribution, indicating that IL provides a better fit for the considered data set.

4.2. Parameter estimation based on the data set

Parameter estimation process is done with the help of R version 4.1.1. We compare the parameter estimation values of KGIL distribution with IL distribution.

Table 4. Comparation of parameter estimation result between IL and KGIL distribution.

Distribution	Parameter			
	$\hat{p}$	$\hat{q}$	$\hat{\lambda}$	$\hat{\beta}$
Inverse Lomax	-	-	3.5542	42,0628
KGIL	0.247558	2.963586	9.631189	171.897472

4.3. Other quantities of the KGIL distribution based on the data set

4.3.1. Cumulative distribution function (cdf) of data set. From Figure 3, we can see the CDF of data set considered based on the KGIL distribution with the estimated parameter values from Table 4.

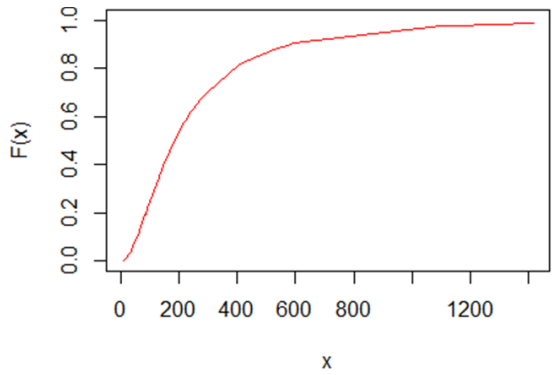

Figure 3. Cumulative distribution function of head-and-neck cancer patient survival time based on KGIL distribution.

4.3.2. Probability density function (pdf) of data set. From Figure 4, we can see the PDF of data set considered based on the KGIL distribution with the estimated parameter values from Table 4. The PDF of the data set is a non-monotone curve.

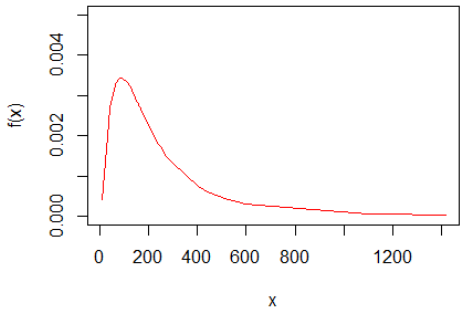

Figure 4. Probability density function (PDF) of head-and-neck cancer survival time using KGIL distribution.

4.3.3. Survival function of data set. From Figure 5, we can see the survival function of data set considered based on the KGIL distribution with the estimated parameter values from Table 4.

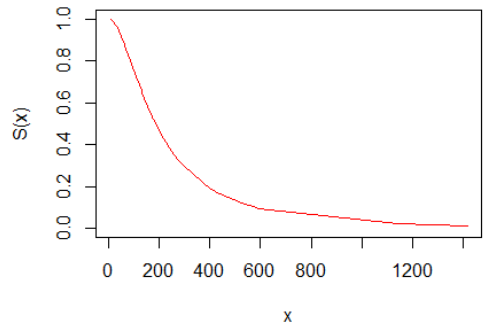

Figure 5. Survival function of head-and-neck cancer survival time using KGIL distribution

4.3.4. Hazard function of data set. From Figure 6, we can see that the data set has a inverted-bathtub hazard shape. The hazard rate constantly increases at the beginning of time, reaches a peak at some point, and slowly decreases for the rest of the observation period.

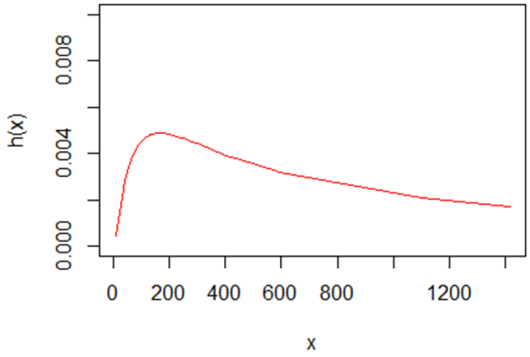

Figure 6. Hazard function of head-and-neck cancer survival time using KGIL distribution

5. Conclusion

KGIL distribution is a newly developed distribution formulated to improve the modelling potential of a data set. It has four parameters, three of it are shape parameters, and one is scale parameter. The main quantities of the distribution can be derived, such as cumulative distribution function, probability density function, survival function, and hazard function. In the Chapter 4, there are three methods used to test whether KGIL distribution can depict the data of head-and-neck cancer patients better than IL. Based on the AIC and BIC values, KGIL is claimed to be the best distribution to model the considered head-and-neck cancer patient data set as it has smaller AIC and BIC value. However, the difference between BIC and AIC values of KGIL and IL are too small. Therefore, we performed a likelihood ratio test (LRT) and it was found that IL is considered to be a good distribution to model the considered data set. From these tests, both distributions are good to model the considered data set. In addition to that, KGIL distribution can address non-monotonicity problem, i.e., inverted-bathtub hazard shape.

References

[1] Tacq J 2010 The Normal Distribution and its Applications *International Encyclopedia Of Education*, pp. 467-473.
[2] Hogg R V, McKean J W, and Craig A T 2019 *Introduction to Mathematical Statistics* (8th ed.). Boston: Pearson p 254.
[3] Kumaraswamy P 1980 A Generalized Probability Density Function for Double-Bounded Random Processes. *Journal of Hydrology*, **46** pp 79-88.
[4] Cordeiro G M and De Castro M 2010 A New Family of Generalized Distributions. *Journal of Statistical Computation and Simulation*, **81** pp 883-898.
[5] Yadav A S, Singh S K, and Singh U 2016 On Hybrid Censored Inverse Lomax Distribution: Application to the Survival Data. *STATISTICA*, **76** pp 185-203.

[6] Ogunde A A, Chukwu A U, and Osegale I O 2023 The Kumaraswamy Generalized Inverse Lomax Distribution and Applications to Reliability and Survival Data. *Scientific African*, **19** pp 1-15.
[7] Efron B 1988 Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve. *Journal of the American Statistical Association*, **83** pp 414-425.

Acknowledgments
The author is grateful for the support and constructive suggestions provided by the supervisors, which improved the paper.