IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# A COMPARISON OF PARAMETRIC METHODS OF REGRESSION MODELS USING CHRONIC KIDNEY DISEASE DATA

[1]Srinivasulu V, [2]Venkataramanaiah M, [3]Ahammad Basha Shaik

[1]Research Scholar, [2]Rtd. Professor, Department of Statistics, Sri Venkateswara University, Tirupati, Andhra Pradesh, India.

[3]Assistant Professor of Statistics, Department of Community Medicine, Narayana Medical College, Nellore, Andhra Pradesh, India.

***Abstract:*** A major global public health issue linked to increased morbidity and mortality is chronic kidney disease (CKD), a continuum of kidney disease ranging from minor kidney impairment to end-stage renal disease (ESRD). For ESRD patients, longevity and quality of life have improved thanks to advancements in dialysis treatments and the availability of kidney transplantation. In survival analysis, survival data might include patient characteristics that are connected to response, survival, and the development of a disease, as well as survival time and response to a particular treatment. Comparing the survival distributions of experimental animals or human patients, predicting the likelihood of response, survival, or mean lifespan, and identifying risk and/or prognostic factors related to response, survival, and the onset of a disease have been the main goals of the study of survival data. In the modelling of survival data under various types of diseases, parametric regression models are frequently utilised. This study uses the CKD data to examine the effectiveness of the four most used parametric regression modelling techniques: Exponential, Weibull, Gamma, and Lognormal. Based on the results and the minimum values of AIC and BIC, it is concluded that the Gamma model performs better than other models.

***Index Terms*** **- Non-communicable disease, Akaike's Information Criterion (AIC), Bayesian information criterion (BIC), Chronic Kidney Disease, Survival analysis, Parametric regression model.**

## I. INTRODUCTION

As per the National Kidney Foundation (NKF), the definition of chronic kidney disease is defined as kidney damage for three months with eGFR < 60 ml/min/1.73 m2 and or structural or functional abnormalities in pathology, imaging, urine analysis, blood composition indicating abnormal kidney function. [1]

Chronic Kidney Disease (CKD), a continuum of kidney disease ranging from mild kidney damage to End Stage Renal Disease (ESRD) is a major public health problem worldwide associated with increased morbidity and mortality.[2] Improvements in techniques of dialysis and availability of renal transplantation have improved survival and quality of life for patients of ESRD. The 5-year survival rate for patients on hemodialysis is 30-50 % in nondiabetics and 25% in diabetics, while the 5-year survival rate for living donor transplantation is 81%. According to the first annual report published by the CKD registry of India, treatment of chronic kidney disease and its advanced stage end stage renal disease is expensive and beyond the reach of average Indian.[2]. Patients with CKD are at high risk for cardiovascular disease (CVD) and cerebrovascular disease, and they are more likely to die of CVD than to develop end-stage renal failure [3]. There is paucity of Indian studies depicting clinical profile of CKD.

Survival analysis can also be used to measure the time to any defined event. Methods for survival analysis allow analysis of such rates without assuming that they are constant. Survival analysis methods are important in trials where participants are entered over a period and have various lengths of follow-up. These methods permit the comparison of the entire survival experience during the follow-up and may be used for the analysis of time to any dichotomous response variable. [4]

Survival analysis involves the time until the occurrence of a specific event can be used to define survival time in general terms. A disease's onset, a patient's response to treatment, a recurrence, or death are all examples of this event. The term "survival time" describes the period of time between the start of the patient's observation and the occurrence of an observed event. It can be measured in years, months, weeks, or days (death). Therefore, the length of remission, the amount of time spent without a tumor, and the period until death can all be considered components of survival time.[5] Data on survival may include length of survival, responsiveness to a particular treatment, and patient traits associated with response, survival, and illness progression. [6]

In the last four decades, survival analysis has emerged as one of the most popular techniques for data analysis across a variety of fields, including criminology, marketing, astronomy, epidemiology, and environmental health. Participants or patients may drop out of clinical and epidemiological research frequently, making them unreachable for follow-up.

Regression models are frequently employed in a variety of disciplines to examine the relationship between an outcome variable and one or more predictor variables. Regression models using parametric techniques are frequently used to model survival data for a variety of diseases, including for the examination of breast cancer survival data, parametric regression models were used to compare five different survival models from the breast cancer registry, nonlinear regression models for heart attack data, and survival models from tuberculosis clinical trials, [7-10] comparison of parametric methods for regression models namely, Exponential, Weibull, Gamma, Lognormal and Log-logistic using the heart attack data.[11-12]

Inspected by the work done in this direction, an attempt is made in the present study to compare the performance of the commonly used parametric regression models in survival analysis namely, Exponential, Weibull, Gamma, and Lognormal distributions using the chronic kidney disease patient's data.

## MATERIALS AND METHODS: -

A retrospective study of 100 chronic kidney disease patterns (eGFR < 60) over a period of one year whose age was 60 years and above was collected from the hospital records of the nephrology department at a tertiary care hospital, Nellore district, AP, India.

The parametric methods for regression model namely, Exponential, Weibull, Gamma, and, Lognormal models are as follows.

### Exponential distribution: -

The exponential distribution is often referred to as a purely random failure pattern. It is famous for its unique "lack of memory". The exponential distribution is characterized by a constant hazard rate λ, its only parameter. When the survival time T follows the Exponential distribution with a parameter λ, the hazard, density and survivorship functions is defined as,

$h(t) = \lambda$; $f(t) = \lambda e^{-\lambda t}$; $S(t) = e^{-\lambda t}$; where $\lambda > 0$

The exponential distribution successfully used as the model for survival time in a study of new anticancer drugs in the L1210 animal leukaemia system [12].

### Weibull distribution: -

The Weibull distribution is a generalization of the exponential distribution. The distribution was proposed by Weibull in the year 1939 and is used in many studies of reliability and human disease mortality, since it allows the survival distribution of a population with increasing, decreasing, or constant risk [13]. A random variable T has the Weibull distribution with the following hazard, density and survivorship functions is defined as,

$h(t) = \lambda\gamma(\lambda t)^{-1}$ ;

$f(t) = \lambda\gamma\ (\lambda t)^{\gamma-1}\ e^{-(\lambda t)^{\gamma}}$;

$S(t) = e^{-(\lambda t)^{\gamma}}$ ; where $\lambda > 0$, $\gamma > 0$

Weibull distribution is preferred for performing survival data analysis in industrial engineering uses. The Weibull distribution was applied to a two-group experiment on vaginal cancer in rats exposed to the carcinogen DMBA [14]. The distribution of the survival period of childhood leukaemia patients was analysed using Weibull distribution [15].

**Gamma Distribution: -**

The gamma distribution, which includes the exponential and chi-square distribution, was used to describe the life of glass tumblers circulating in a cafeteria [16] and as a statistical model for life length of materials and an application of the gamma distribution to the lifetime of aluminium coupon [17]. Stored platelet survival data analysis by a gamma model [18]. This distribution has been used frequently as a model for industrial reliability problems and human survival. The probability density function, hazard function and survivorship function of a gamma distribution is defined as follows,

$$f(t) = \frac{\lambda}{\Gamma(\gamma)} (\lambda t)^{\gamma-1} e^{-\lambda t};$$

$$h(t) = \frac{\lambda(\lambda t)^{n-1}}{(n-1)! \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!}};$$

$$S(t) = e^{-t} \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} \text{ ; where } \gamma>0.$$

Where γ is the shape parameter and ɼ is the gamma function.

The hazard function of a gamma distribution can provide varieties of forms depending on the value of the ã parameter.

**Lognormal distribution: -**

The lognormal distribution is defined as the distribution of a variable whose logarithm follows the normal distribution. The theory of the lognormal distribution was described by McAlister in the year 1879 and is used in many areas of medicine. A random variable T has the lognormal distribution with the following density, hazard and survivorship function is,

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log t-\mu}{\sigma}\right)^2};$$

$$h(t) = \frac{\frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log t-\mu}{\sigma}\right)^2}}{1-G\left(\log \frac{at}{\sigma}\right)} \text{ ; where } \sigma > 0.$$

$$S(t) = 1 - G\left(\log \frac{at}{\sigma}\right);$$

Where G(y) is the cumulative distribution function.

A review of lognormal distribution and its application in biology, followed by applications in cancer research [19]. Its history, properties, estimation problems, and uses in economics have been discussed [20]. The distribution of survival time of several diseases such as Alzheimer's disease, Hodgkin's disease and chronic leukaemia could be rather closely approximated by a lognormal distribution since they are markedly skewed to the right and the logarithms of survival times are approximately normally distributed. In a study of chronic lymphocytic and myelocytic leukaemia patients, applied the lognormal distribution to analyse survival data of 649 white residents of Brooklyn diagnosed from 1943 to 1952 [21]. The lognormal distribution is suitable for survival patterns with an initially increasing and then decreasing hazard rate.

**Model selection criteria**

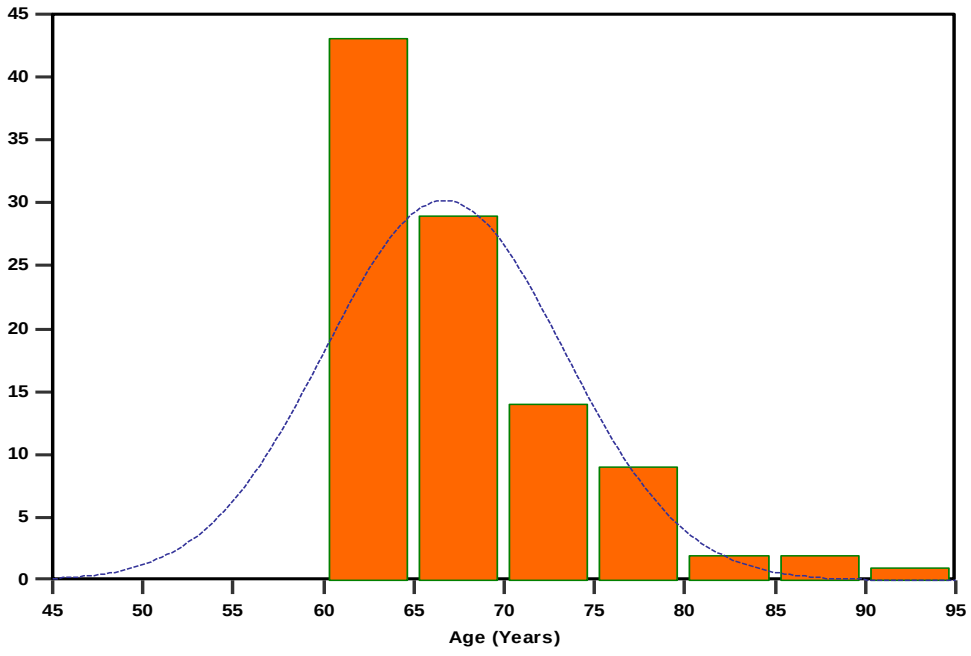
When comparing the efficacy of the models related to the CKD data and to compare the goodness of fit of the fitted parametric models in terms of fitting the observed data, log likelihood (LL), the Akaike's Information Criterion (AIC) (Akaike, 1973) or Bayesian information criterion (BIC) (Schwarz, 1978) can be used for model selections [24]. A lower value of AIC indicates a better model. In the case of analysing clinical trial data, both BIC and AIC will give similar results as the sample sizes are relatively small in clinical trials (Lindsey and Jones, 1998). The formula for AIC equation is defined as, AIC = −2LL + (2c+a), where LL is the log likelihood statistic, 'c' indicates number if parameters in the survival distribution function and 'a' denotes the number of parameters in the model. The statistical analysis has been done by using statistical software STATA Version 17 (STATA Inc., NC, USA). All the p values having less than 0.05 were considered as statistically significant.

## RESULTS AND DISCUSSION

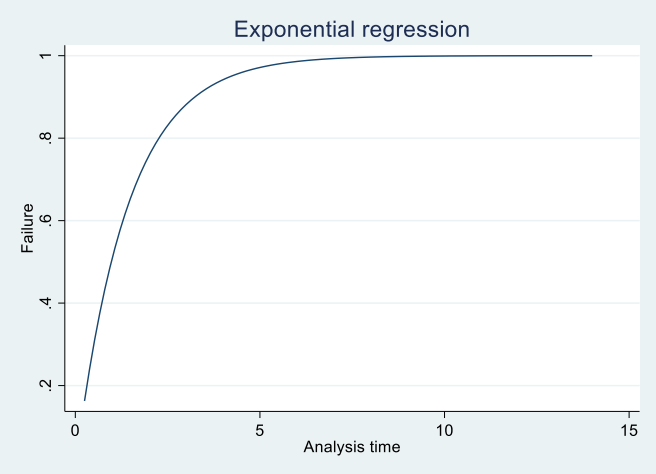
In a total of 100 CKD patients data, 57.0% of patients were males, and 43.0% of patients were females. The mean age of patients was 66.69±6.61 years with a range of age was 60 years to 90 years respectively. The mean of BMI was 21.51±3.03. The descriptive statistics for CKD patient's data was shown in Table-1 and the histogram of age distribution (years) was shown in Figure-1.

**Table-1: Descriptive Statistics for the CKD patient's data**

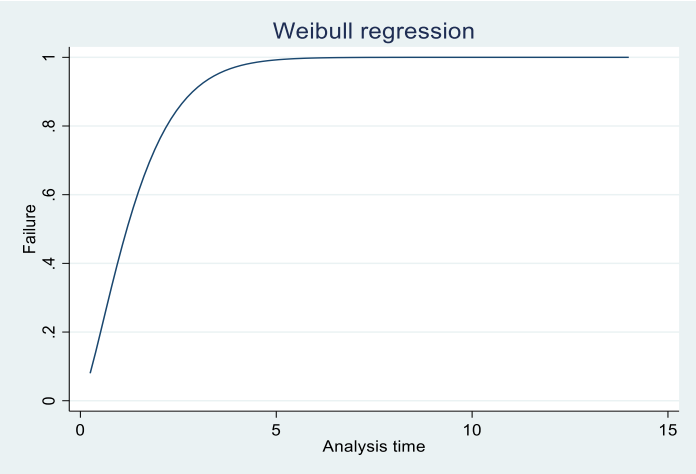
| Variable (s) | Mean | 95% Con. Limits of Mean | | Std. Dev. |
|---|---|---|---|---|
| | | Lower Limit | Upper Limit | |
| Age | 66.690 | 65.38 | 68.00 | 6.6084 |
| BMI | 21.506 | 20.91 | 22.11 | 3.0307 |
| MCV | 82.547 | 80.97 | 84.13 | 7.9733 |
| SGOT | 24.540 | 20.90 | 28.18 | 18.3427 |
| SGPT | 24.100 | 20.63 | 27.57 | 17.4775 |
| Creatinine | 4.516 | 3.72 | 5.31 | 4.0172 |
| eGFR | 19.630 | 16.99 | 22.27 | 13.3251 |
| pCO2 | 27.119 | 25.43 | 28.81 | 8.5165 |
| Sr_Sodium | 135.630 | 134.24 | 137.02 | 6.9843 |
| Sr_Potassium | 4.568 | 4.36 | 4.78 | 1.0553 |
| EF | 52.460 | 50.22 | 54.70 | 11.2703 |



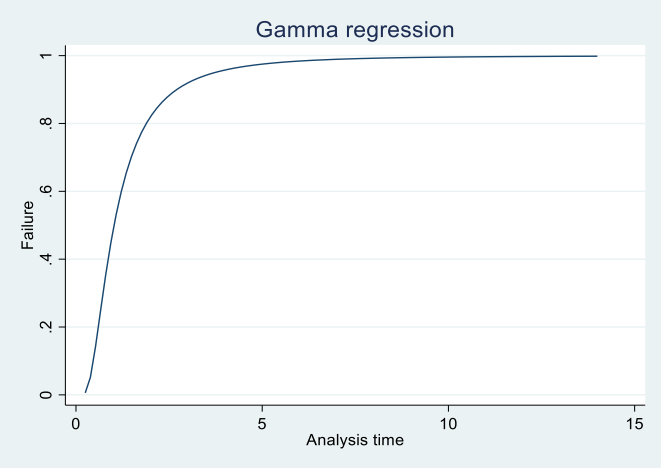

**Graph-1: Histogram of age distribution (Years) of CKD data**



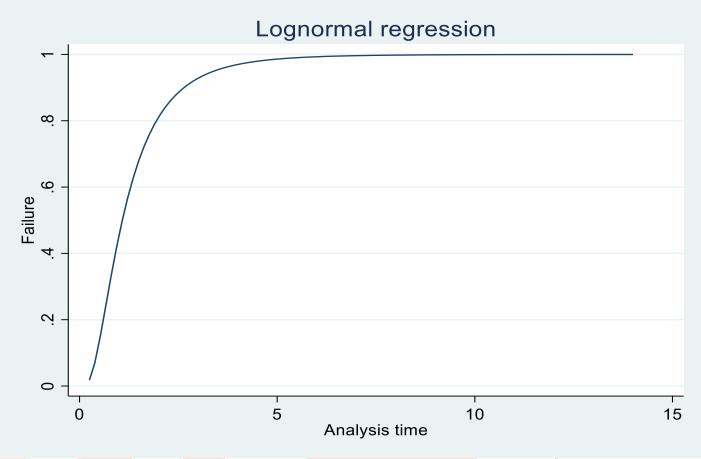

(a)




(b)




(c)




(d)

**Figure-2: Graphical presentation of failure rate functions of (a) Exponential, (b) Weibull, (c) Gamma and (d) Lognormal Distributions for CKD data.**

**Table 2: Covariates comparison of analysis of maximum likelihood parameter estimates of five parametric methods for Regression models to CKD data.**

| Parameter | Exponential | Weibull | Gamma | Lognormal |
|---|---|---|---|---|
| Age | -0.014 (0.017) | -0.017 (0.013) | -0.003 (0.010) | -0.006 (0.010) |
| BMI | -0.026 (0.032) | -0.026 (0.024) | -0.037 (0.025) | -0.033 (0.024) |
| MCV | -0.023 (0.013) | -0.026 (0.010) * | -0.007 (0.008) | -0.012 (0.008) |
| SGOT | 0.011 (0.010) | 0.120 (0.008) | 0.006 (0.005) | 0.008 (0.060) |
| SGPT | -0.013 (0.044) | -0.014 (0.008) | -0.011 (0.006) | -0.110 (0.006) |
| Creatinine | 0.032 (0.043) | 0.042 (0.036) | 0.017 (0.022) | 0.017 (0.024) |
| eGFR | -0.003 (0.012) | -0.002 (0.010) | 0.005 (0.008) | 0.002 (0.008) |
| pCO2 | -0.003 (0.015) | 0.001 (0.012) | -0.010 (0.008) | -0.004 (0.009) |
| Sr. Sodium | 0.019 (0.011) | 0.022 (0.008) * | 0.001 (0.008) | 0.111 (0.008) |
| Sr. Potassium | 0.114 (0.106) | 0.111 (0.081) | 0.085 (0.070) | 0.099 (0.073) |
| EF | 0.010 (0.010) | 0.125 (0.008) | 0.004 (0.006) | 0.006 (0.007) |

***Values are expressed as maximum likelihood estimate (standard error), * - Significant.***

**Table 3: Model selection criteria of the fitted models for CKD data**

| Parametric distribution | -2LL | AIC | BIC |
|---|---|---|---|
| Exponential | -126.411 | 274.822 | 303.479 |
| Weibull | -118.657 | 261.315 | 292.577 |
| Gamma | -103.677 | 233.353 | 267.220 |
| Lognormal | -105.762 | 235.525 | 266.787 |

Table-2 shows that the Covariates comparison of analysis of maximum likelihood parameter estimates of five parametric methods for Regression models to CKD data and Figure-2 showed Graphical presentation of failure rate functions of Exponential, Weibull, Gamma, and Lognormal Distributions for CKD data. Table 3 shows that the values of -2LL, AIC and BIC criteria for the fitted models. The AIC and BIC results provide strong evidence that Gamma model is performing better than other models.

In a study, the results indicated that the early detection of a cancer at a young patient age and in primary stages is important to increase survival from gastric cancer [7]. In a study result, the Gompertz distribution model is determined as the most suitable model than the Weibull, Lognormal and Gamma models for the breast cancer data by considering lower value of AIC and as a result of the analyses depending on the parameter estimates, age variable was not found as a risk factor [8]. The Gamma model was the best fitting parametric model for applied likelihood-based criteria in a tuberculosis clinical trial data [10]. The lognormal model is better than the Exponential, Weibull, Gompertz, Lognormal and Log-logistic models in the analysis of breast cancer data by using likelihood ratio values [11]. A study showed that, likelihood-based criteria for model selection indicated that the Weibull model was the best fitting parametric model for predicting survival following both HIV and AIDS diagnoses [23]. In this study, the results stated that the Gamma distribution model is most suitable than the Exponential, Weibull, and Lognormal models for CKD data. A parametric survival model with restricted cubic spline function was applied to assess prognostic factors was done by Jha et al. (2018) [2], they concluded that the risk of death was dramatically increased after developing kidney failure, and they identified three major transitions for prognostic factors that effects the CKD progression such as Diabetes, hypertension, and cardiovascular disease that increased the risks of death before and after kidney failure. In a study of Shaik et al. (2015) [10], compared the parametric methods for regression models namely, Exponential, Weibull, Gamma, Lognormal and Log-logistic using the heart attack data, they concluded that Weibull and Gamma models are performing better than other models. Our study is coinciding with this study.

## CONCLUSIONS: -

Many researchers, who used various approaches in their tests, have proposed and advised the various models. The results of comparing various models revealed that the Gamma model is the most effective, outperforming all others. The results need to be confirmed by other research.

## REFERENCES


[1] National Kidney Foundation (2002). K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. American journal of kidney diseases: the official journal of the National Kidney Foundation, 39(2 Suppl 1), S1-S266.

[2] Jha VK and Shashibhushan. Clinical Profile of Chronic Kidney Disease Patients in a Tertiary Care Hospital-An Observational Study. J Nephrol Kidney Dis. 2018; 2(2): 1016. https://dx.doi.org/10.36876/smjnkd.1016.

[3] Schieppati A, Giuseppe R. Chronic renal diseases as a public health problem: Epidemiology, Social and economic implications. Kidney Int. 2005; 68:s7-s10.

[4] Nelson W (1982). Applied Life Data Analysis. Wiley, New York.

[5] Lee, E.T., Statistical methods for survival data analysis, 3rd Edition, John Willey &Sons, NY, (1992).

[6] Lee, E.T., Go T, Oscar (1997). Survival Analysis in Public Health Research, Annu. Rev. Public Health, 18:105–34.

[7] Ahmad Reza Baghestani, Ebrahim Hagizadeh, and Seyed Reza Fatemi., Parametric model to analyse the survival of gastric cancer in the presence of interval censoring. Tumor, 69, 433-7, (2010).

[8] Vallinayagam V, Prathap, S, Venkatesan, P. Non-Linear Regression Models for Heart Attack Data – An Empirical Comparison. Indian Journal of Applied Research, 2014;4(6):332-334.

[9] Zelen, M., Applications of Exponential Models to Problems in Cancer Research. Journal of the Royal Statistical Society, Series A, 1966;129:368-398.
[10] Ahammad Basha Shaik Venkataramanaiah M Madhusudan G (2015). Comparison of parametric methods for regression model by using cardiovascular disease survival data. International Journal of Scientific and Innovative Mathematical Research, 3,2:457-462.
[11] Hosmer, D.W. and Lemeshow, S. and May, S, Applied Survival Analysis: Regression Modelling of Time to Event Data: 2nd Ed., John Wiley and Sons Inc., New York, (2008).
[12] Zelen, M., Applications of Exponential Models to Problems in Cancer Research. Journal of the Royal Statistical Society, Series A, 129, 368-398 (1966).
[13] Collett, D, Modelling Survival data in Medical Research. Chapman & Hall,London, (2003).
[14] Pike, M. C., A Method of Analysis of a Certain Class of Experiments in Carcinogenesis. Biometrics, 22, 142-161, (1966).
[15] Viscomi S, Pastore G, Dama E, et al, Life expectancy as an indicator of outcome in follow-up of population-based cancer registries: the example of childhood leukaemia, Ann. Oncol, 17(1), 167-71, (2006).
[16] Fd Brown, G. W., and Flood, M. M., Tumbler Mortality. Journal of the American Statistical Association, 42, 562-574, (1947).
[17] Birnbaum, Z. W., and Saunders, S. C., A Statistical Model for Life-Length of Materials. Journal of the American Statistical Association, 53, 151-160, (1958).
[18] Bolin RB, Greene JR. Stored platelet survival data analysis by a gamma model. Transfusion, 26(1):28-30, (1986).
[19] Boag, J. W., Maximum Likelihood Estimates of Proportion of Patients Cured by Cancer Therapy. Journal of the Royal Statistical Society, Series B, 11, 15, (1949).
[20] Aitchison, J., and Brown, J. A. C., The Lognormal Distribution. Cambridge University Press, Cambridge, (1957).
[21] Feinleib M and Macmahon B, Variation in the duration of survival of patients with the chronic leukemia's, Blood, 15(3), 332-49, (1960).
[22] 24. Burnham KP, Anderson DR. Multi model inference: understanding AIC and BIC in model selection. Sociol Meth Res.,33:261–304, (2004).
[23] Nakhaee, F. and Law, M. (2011): Parametric modelling of survival following HIV and AIDS in the era of highly active antiretroviral therapy: data from Australia. Eastern Mediterranean Health Journal, 7(3).
[24] Erohildes Ferreira R, Sanders-Pinheiro H and Basile Colugnati FA (2023) A proposal to analyze the progression of non-dialytic chronic kidney disease by surrogate endpoints: introducing parametric survival models. Front. Med. 10:1029165. doi:10.3389/fmed.2023.1029165.