Regression Analysis of Exponentiated Gumbel Type-2 Life time Model: Bayesian Approach

A. A. Ogunde[1]; A. U. Chukwu[1]; J. Audu[2]; B. Ajayi[3]

[1]Department of Statistics,
University of Inadan, Nigeria.

[2]Department of Statistics,
Federal School of Statistics, Ibadan, Nigeria.

[3]Department of Mathematics and Statistics,
Federal Polytechnic Ado-Ekiti, Nigeria.
E-mail: debiz95@yahoo.com[1]

***Abstract* — This study is concerned with the estimation of shape parameters of Exponentiated Gumbel type-2 distribution using various Bayesian approximation techniques. Regression analyses were carried out for real survival data problems with random censoring mechanisms. Different informative and non-informative priors were used to obtain the Bayes' estimate of parameter of Exponentiated Gumbel type-2 distribution under different approximation techniques. For comparing the efficiency of the obtained results, a simulation study was carried out using R-software.**

Keywords - *Non-informative priors, Exponentiated Gumbel type-2 distribution, Bayesian approximation, R-software, random censoring.*

I. Introduction

One way of handling heterogeneity in a population is by incorporating regressor variables in the model. When handling lifetime data, oftentimes, we incorporate regressor variables in other to boost the performance of the model in relation to the characteristics (concomitant variables) that that affect the response variable (survival time). For example when carrying out a study on survival time for lung cancer patients; factors such as age of patient, the types of tumor, the time of the first diagnosis etc. can be relevant factors to be considered.

Regression model with lifetimes as the response variable and the concomitant variables as the regressor variables gives room for such additional factors to be introduced in a statistical analysis.

One important feature of survival function is that the presence of censoring creates problems in the analysis. Life time data are censored when the specific time of failure for a given trial is unknown. When analysing censored data, Bayesian method has a unique advantage over the classical method. From a classical point of view, confidence interval and other inferential statements must be made with respect to repeated sampling of the data an advantage of the Bayesian approach is that only the censoring pattern; e.g. a right censored failure time, is relevant, not the type of censoring scheme, such as Type 1, Type II or random sampling that produced it.

II. Research Methodology

2.0 The Exponentiated Gumbel Type-2 distribution

The cumulative density function (cdf) of Exponentiated Gumbel Type-2 (EGT-2) distribution as given by Okorie et al. (2017) is given as

$$F(t;\alpha,\lambda,\phi) = 1-\left(1-e^{-\lambda t^{-\alpha}}\right)^{\phi}, \quad t>0;\ \alpha,\lambda,\phi>0 \qquad (1)$$

And the associated probability density function is given by

$$f(t;\alpha,\lambda,\phi) = \alpha\lambda\phi t^{-\alpha}e^{-\lambda t^{-\alpha}}\left(1-e^{-\lambda t^{-\alpha}}\right)^{\phi-1}, \quad t>0;\ \alpha,\lambda,\phi>0 \qquad (2)$$

where α and ϕ are the shape parameters and λ is the scale parameter. The graph of the pdf for EGT-2 distribution is given by Fig 1.0 for various values of the parameters.

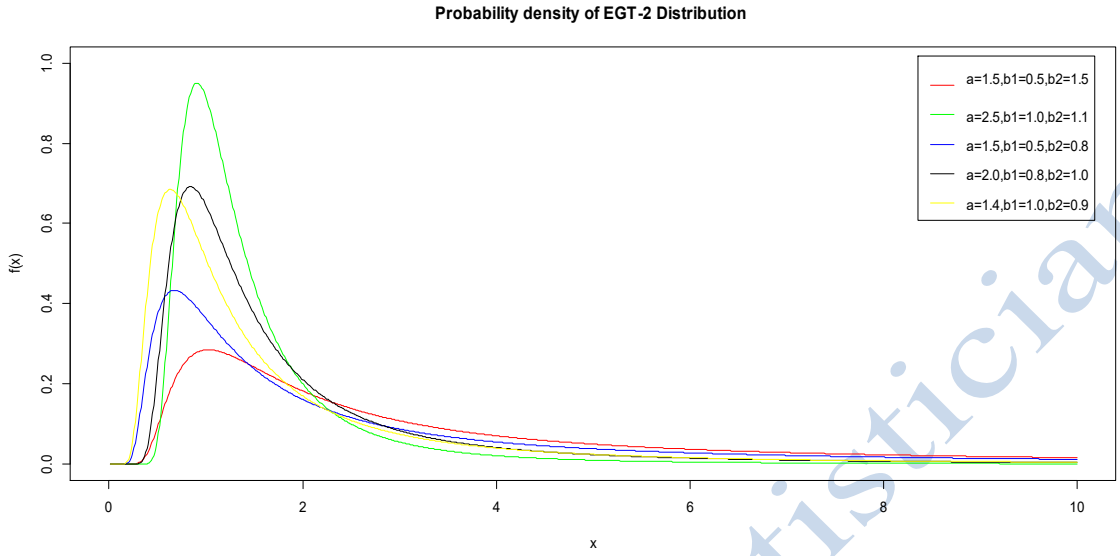

Figure 1.0: A graph of the pdf of EGT-2 distribution

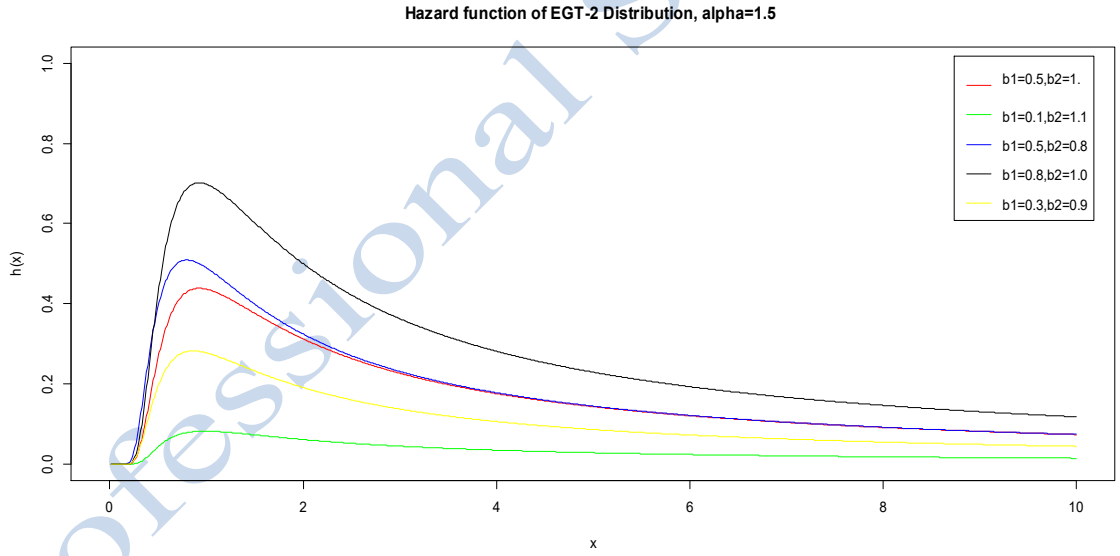

Figure 2.0: A graph of the hazard function of EGT-2 distribution

An expression for the survival and the hazard function is given below in equations (3) and (4) respectively;

$$S(t,\alpha,\lambda,\phi) = \left(1 - e^{-\lambda t^{-\alpha}}\right)^{\phi}, \qquad t > 0;\ \alpha,\lambda,\phi > 0 \qquad (3)$$

$$h(t;\alpha,\lambda,\phi) = \frac{\alpha\lambda\phi t^{-\alpha} e^{-\lambda t^{-\alpha}} \left(1 - e^{-\lambda t^{-\alpha}}\right)^{\phi-1}}{(1 - e^{-\lambda t^{-\alpha}})^{\phi}} \qquad (4)$$

The graph of the survival and the hazard function are given by Fig 2.0.

2.1 Construction of Exponentiated Gumbel Type-2 regression Model

Tablemann and Kim (2004) model distribution using the hazard function. Assuming the hazard function at time t for an individual has the form

$$h(t/x) = h_0(t).e^{x^T\beta} \quad (5)$$

Therefore for EGT-2 distribution, we have

$$h(t/x) = \frac{\alpha\lambda\phi t^{-\alpha}e^{-\lambda t^{-\alpha}}\left(1-e^{-\lambda t^{-\alpha}}\right)^{\phi-1}}{(1-e^{-\lambda t^{-\alpha}})^{\phi}}e^{x^T\beta} \quad (6)$$

Where, β=$[\beta_1, \beta_2, \ldots, \beta_p]$ is a vector of regression parameters. The function $h_0(t)$ is known as the baseline hazard. It is define as the value of the hazard function when the covariate vector $x = 0$ *or* $\beta = 0$. The equation (6) implies that the covariate act multiplicatively on the hazard rate.

The survival function of T is given by

$$S(t/x) = \exp(-h(t/x)t)$$

$$= exp\left\{-\frac{\alpha\lambda\phi t^{-\alpha}e^{-\lambda t^{-\alpha}}\left(1-e^{-\lambda t^{-\alpha}}\right)^{\phi-1}}{(1-e^{-\lambda t^{-\alpha}})^{\phi}}e^{X^T\beta}\right\} \quad (7)$$

Thus, the pdf of T given x is

$$f(t/x) = h(t/x).S(t/x) \quad (8)$$

Therefore, putting equation (8) into (7), we have an expression for the pdf given by

$$f(t/x) = exp\left\{-\frac{\alpha\lambda\phi t^{-\alpha}e^{-\lambda t^{-\alpha}}\left(1-e^{-\lambda t^{-\alpha}}\right)^{\phi-1}}{(1-e^{-\lambda t^{-\alpha}})^{\phi}}e^{X^T\beta}\right\}\times$$

$$\frac{\alpha\lambda\phi t^{-\alpha}e^{-\lambda t^{-\alpha}}\left(1-e^{-\lambda t^{-\alpha}}\right)^{\phi-1}}{\left(1-e^{-\lambda t^{-\alpha}}\right)^{\phi}}e^{X^T\beta} \quad (9)$$

III. Liklihood Functions and Priors Elicitation

3.0 Likelihood function of EGT-2 regression model with Censoring

Suppose that there are n subjects under study, and that t_i *and* t_{ci} are respectively the associated survival time and censoring time respectively. The $t's$ are assumed to be independent and identically distributed with density $f(t)$ and survival time $S(t)$. The exact survival time t_i of an individual will be observed only if, $t_i \leq t_{ci}$. The framework for the set oif data for n pair of random variables (y_i, δ_i), where

$$y_i = min(t_i, t_{ci})$$

and

$$\delta_i = \begin{cases} 1 & if\ t_i \leq t_{ci} \\ 0 & if\ t_i > t_{ci} \end{cases} \quad (10)$$

Then the likelihood function for $(\beta, h_0(t))$ for a set of right censored data on n subjects is given by

$$L \propto \prod_{i=1}^{n} f(y_i/x_i)^{\delta_i} S(t_{ci}/x_i)^{1-\delta_i} \quad (11)$$

Putting equation (7) and (9) in (11) , we have

$$L = \prod_{i=1}^{n}\left(exp\left\{-\frac{\alpha\lambda\phi y_i{}^{-\alpha}e^{-\lambda y_i{}^{-\alpha}}\left(1-e^{-\lambda y_i{}^{-\alpha}}\right)^{\phi-1}}{(1-e^{-\lambda y_i{}^{-\alpha}})^{\phi}}e^{x_i{}^T\beta}\right\}\times \frac{\alpha\lambda\phi y_i{}^{-\alpha}e^{-\lambda y_i{}^{-\alpha}}\left(1-e^{-\lambda y_i{}^{-\alpha}}\right)^{\phi-1}}{(1-e^{-\lambda y_i{}^{-\alpha}})^{\phi}}e^{x_i{}^T\beta}\right)^{\delta_i}$$

$$\times\left(exp\left\{-\frac{\alpha\lambda\phi t_{ci}{}^{-\alpha}e^{-\lambda t_{ci}{}^{-\alpha}}\left(1-e^{-\lambda t_{ci}{}^{-\alpha}}\right)^{\phi-1}}{\left(1-e^{-\lambda t_{ci}{}^{-\alpha}}\right)^{\phi}}e^{x^T\beta}\right\}\right)^{1-\delta_i}$$

3.1 Prior

Here we set the prior to the EGT-2 proportional hazard model whose likelihood function is given in equation (13). If $Y \sim EGT-2(\alpha, \lambda, \phi)$. Prior probabilities are specified for α, ϕ and β

$$\alpha \sim half\ Cauchy\ (\gamma)$$
$$\phi \sim half\ Cauchy\ (\gamma)$$

$$p(\alpha/\gamma) = \frac{2\gamma}{\pi(\alpha^2+\gamma^2)},\ \alpha > 0 \quad (13)$$

$$p(\phi/\gamma) = \frac{2\gamma}{\pi(\phi^2+\gamma^2)},\ \phi > 0 \quad (14)$$

The half Cauchy distribution with scale parameter $\gamma = 25$ is used as non-informative prior distribution for shape parameter. As Gelman and Hill (2007) recommend that, the uniform or if more information is necessary the half-Cauchy distribution is almost flat.

Since, $\lambda > 0$ and β can take any value on the real line, hence we consider the log link function.

$$log(\lambda) = X^T\beta \quad (15)$$
$$\lambda = e^{X^T\beta}$$

where, X is the model matrix and β is the vector of the regression coefficients.

Each component of the β parameters is assigned a weak informative Gaussian prior probability distribution. Assuming that β_i's are independently distributed as normal with mean=0 and standard deviation=1000, in such that a flat prior can be observed. The large variance indicates a lot of uncertainty about each β, and hence, it can be regarded as a weak informative distribution.

$$\beta_j \sim N(0,1000).$$

Then the joint posterior distribution is given by

$$p(\beta, \alpha, \phi/y, X)$$
$$= p(y/\alpha, \phi, \beta, X).p(\alpha).p(\beta).p(\phi) \quad (16)$$
$$p(\beta, \alpha, \phi/y) \sim \prod_{i=0}^{n} f(t/x)^{\delta} S(t_i/x_i)^{1-\delta_i}.p(\alpha).p(\beta).p(\phi) \quad (17)$$

$$p(\beta,\alpha,\phi/y)\sim\prod_{i=0}^{n}\left(exp\left\{-\frac{\alpha\lambda\phi y_i^{-\alpha}e^{-\lambda y_i^{-\alpha}}\left(1-e^{-\lambda y_i^{-\alpha}}\right)^{\phi-1}}{(1-e^{-\lambda y_i^{-\alpha}})^{\phi}}e^{x_i^T\beta}\right\}.\frac{\alpha\lambda\phi y_i^{-\alpha}e^{-\lambda y_i^{-\alpha}}\left(1-e^{-\lambda y_i^{-\alpha}}\right)^{\phi-1}}{(1-e^{-\lambda y_i^{-\alpha}})^{\phi}}e^{x_i^T\beta}\right)^{\delta_i}$$
$$\times\left(exp\left\{-\frac{\alpha\lambda\phi t_{ci}^{-\alpha}e^{-\lambda t_{ci}^{-\alpha}}\left(1-e^{-\lambda t_{ci}^{-\alpha}}\right)^{\phi-1}}{\left(1-e^{-\lambda t_{ci}^{-\alpha}}\right)^{\phi}}e^{x_i^T\beta}\right\}\right)^{1-\delta_i}\cdot\left\{\frac{2\gamma}{\pi(\alpha^2+\gamma^2)}\right\}.\left\{\frac{2\gamma}{\pi(\alpha^2+\gamma^2)}\right\}$$
$$\times\prod_{j=1}^{p}\left\{\frac{1}{\sqrt{2\pi}1000}e^{-\frac{\beta_j^2}{2*1000^2}}\right\} \quad (18)$$

Marginal for β

$$p(\beta/y,X)=\int_0^{\infty}\int_0^{\infty}p(\beta,\alpha,\phi/y,X)d\alpha d\phi \quad (19)$$

Marginal for α

$$p(\alpha/y,X)=\int_{-\infty}^{\infty}\int_0^{\infty}p(\beta,\alpha,\phi/y,X)d\beta d\phi \quad (20)$$

Marginal for ϕ

$$p(\phi/y,X)=\int_{-\infty}^{\infty}\int_0^{\infty}p(\beta,\alpha/y,X)d\alpha d\beta \quad (21)$$

where β is a vector of length $(p+1)$.

In Bayesian regression analysis close form for posterior distribution of β are generally not available, and therefore there is call to use numerical integration or Markov chain Monte Carlo methods.

IV. APPLICATION

The data was obtained from Krall et al. (1975), related to 48 patients, all of whom were aged between 50 and 80 years. This data had also been discussed by Collett (1994, 2003). Some of this patient had not died by the time the study was completed, and so these individuals contribute right censored survival times.

Table 1.0. Approximated Posterior mean, standard errors and 95% credible region for the multiple myeloma data (n=48) by Laplace Approximation under the assumption of EGT-2 regression Model

	Mode	*SD*	*LB*	*UB*
Intercept	1.9491	0.14309	1.6629	2.2352
Bun	−0.0048	0.00172	−0.0082	−0.0013
Hb	0.0096	0.0198	−0.0300	0.0494
Protein	0.2586	0.1152	0.0282	0.4889
Log.shape1	−1.0879	0.2449	−1.5777	−0.5981
Log.shape2	3.0519	0.8456	1.3608	4.7431

Table 2.0: Approximated posterior summary of regression model of EGT-2 distribution

	Mode	*SD*	*MCSE*	*ESS*	LB	Median	UB
Intercept	1.92	0.143	0.004	1000	1.694	1.921	2.220
Bun	−0.006	0.002	0.000	1000	−0.009	−0.006	−0.002
Hb	0.023	0.026	0.001	1000	−0.029	0.0204	0.068
Protein	0.243	0.118	0.004	1000	−0.037	0.204	0.470
*Log.shape*1	−0.860	0.272	0.009	1000	−0.859	−0.901	−0.193
*Log.shape*2	2.323	0.882	0.028	1000	0.480	2.283	4.362
Deviance	301.15	3.143	0.099	1000	296.05	301.19	308.83
LP	−189.10	1.567	0.050	1000	−0.019	−189.27	−187.35
*Shape*1	0.439	0.123	0.004	1000	2.600	0.423	0.824
*Shape*2	16.071	21.845	0.691	1000	1.615	9.812	78.392

Table 3.0. Simulated posterior summary of regression model of EGT-2 distribution

	Mean	*SD*	*MCSE*	*ESS*	*LB*	*Median*	*UB*
Intercept	1.936	0.083	0.002	1696.860	1.772	1.936	2.098
Burn	−0.005	0.001	0.000	1725.199	−0.007	−0.005	−0.002
Hb	0.001	0.011	0.000	1337.962	−0.012	0010	0.034
Protein	0.261	0.069	0.002	1428.530	0.128	0.261	0.396
*log.shape*1	−1.066	0.144	0.004	1615.552	−1.343	−1.072	−0.775
*Log.shape*2	2.959	0.496	0.001	1353.180	1.988	2.953	3.929
Deviance	296.88	1.308	0.046	1151.439	294.7	296.74	299.78
LP	−187.66	0.603	0.023	1049.572	−189.14	−187.54	−186.82
*Shape*1	0.348	0.051	0.001	1622.436	0.026	0.342	0.461
*Shape*2	21.769	11.197	0.319	1272.029	7.298	19.162	50.868

Table 4.0. Simulated posterior summary of regression model of EGT-2 distribution

	Mean	*SD*	*MCSE*	*ESS*	*LB*	*Median*	*UB*
Intercept	1.936	0.0836	0.002	1696.860	1.772	1.936	2.0981
Bun	−0.005	0.0010	0.000	1725.199	−0.007	−0.005	−0.003
Hb	0.010	0.0117	0.000	1337.962	−0.011	0.010	0.034
Protein	0.261	0.0691	0.002	1428.530	0.128	0.261	0.396
*Log.shape*1	−1.066	0.1447	0.004	1615.552	−1.343	−1.072	−0.775
*Log.shape*2	2.959	0.4955	0.014	1353.180	1.988	2.953	3.929
Deviance	296.81	1.3081	0.046	1151.439	294.73	296.74	299.78
LP	−187.660	0.6032	0.022	1049.572	−189.14	−187.54	−186.82
*Shape*1	0.348	0.0513	0.001	1622.436	0.261	0.342	0.461
*Shape*2	21.769	11.1972	0.319	1272.029	7.298	19.162	50.868

V. Output of Exponentiated Weibull Distribution using Laplace Approximation

The object M1 gives two summaries, Summary1 and summary2 is the summary in the form of posterior mode and modal variance. As the data contain four regressor variables and on the basis of this Bayesian analysis we will have to conclude that which regressor variable is appropriate for modelling survival data.

Table 1 and Table 2 provide the approximated posterior summary of regression model of EGT-2 distribution. The table 3 and Table 4 contain simulated posterior mean, posterior standard deviation and 95% credible region. Table 1 and Table 2 show that only two out of three regressor variables of multiple myeloma patients are significant. The covariates Bun and protein have credible regions (-0.0082, -0.0013) and (0.0282, 0.4889) respectively, which does not include zero and hence they are appropriate regressor variable for modelling survival data. The credible region for age variable is (-0.0300, 0.0494) which includes zero in it, hence is not a significant regressor variable for modelling multiple myeloma data.

Figure 3, figure 4 and figure 5 drawn below is the graph of Trace and posterior density plots for variables of EGT-2 distribution by Independent Metropolitan algorithm.

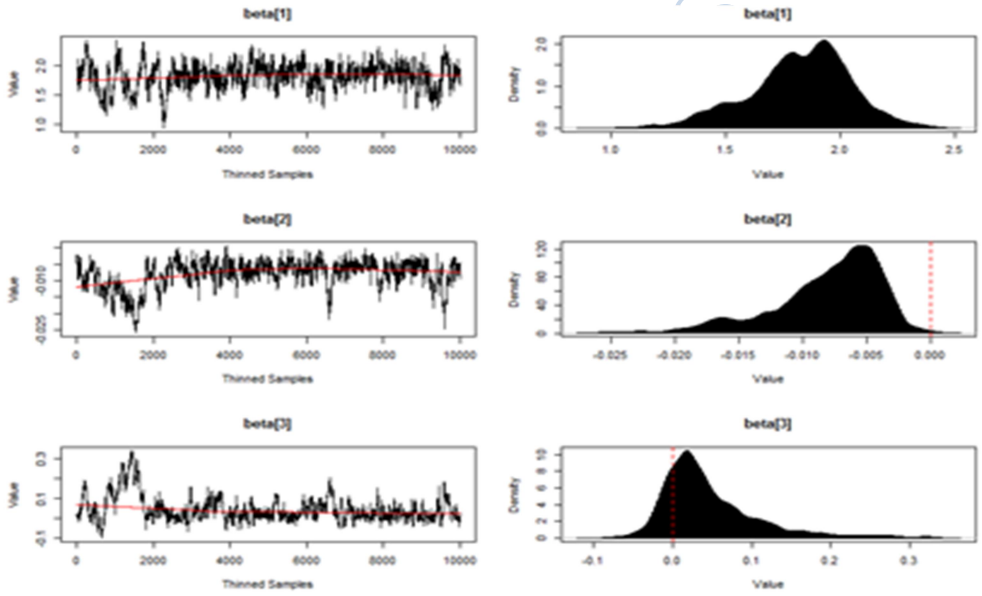

Figure 3: Trace and Posterior density plots for variables of EGT-2 distribution by IM algorithm

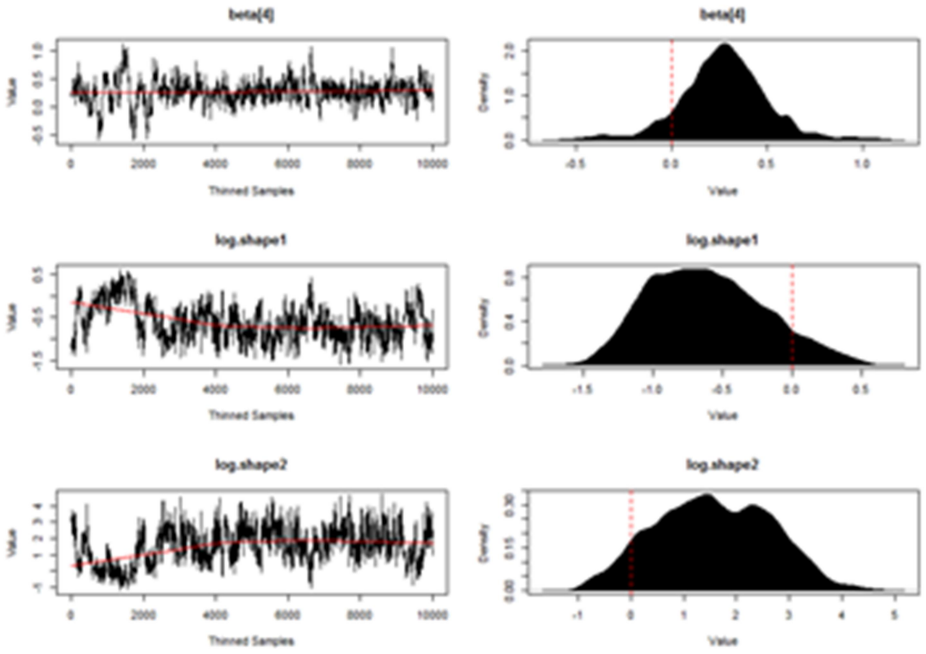

Figure 4.0 Trace and Posterior density plots for variables of EGT-2 distribution by IM algorithm

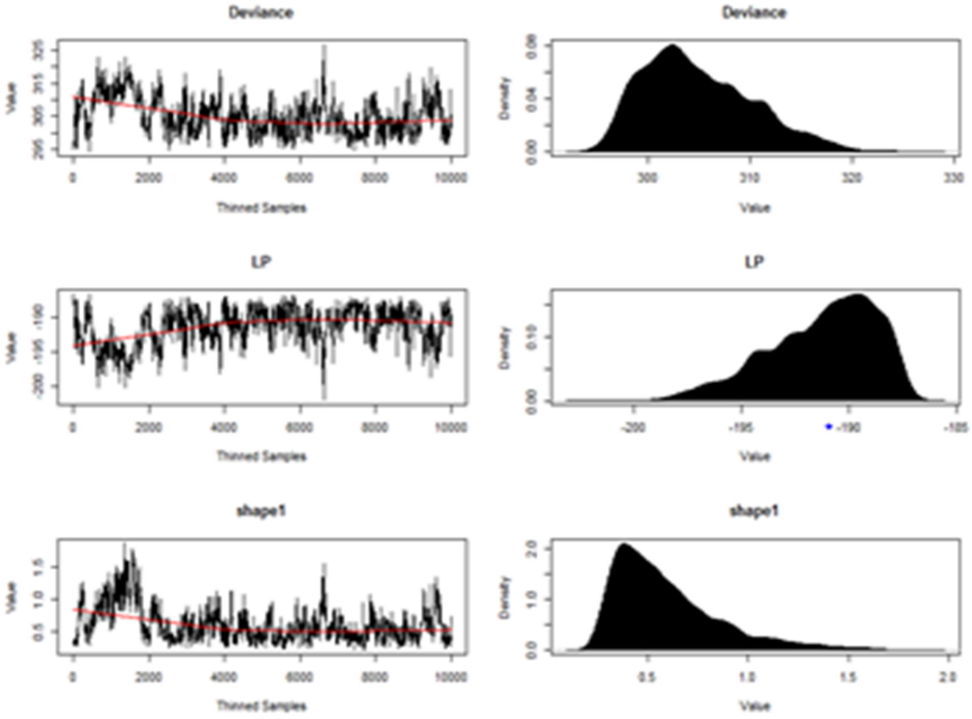

Figure 5.0 Trace and Posterior density plots for variables of EGT-2 distribution by IM algorithm

VI. Conclusion

In this paper, Bayesian approach has been employed to model the real survival data under the assumption Exponentiated Gumbel type-two distribution. These distribution have been used as a Bayesian model to fit the survival data.

This paper includes the derivation of joint and marginal posterior densities of the distribution. Asymptotic approximation and simulated posterior summary of regression model of EGT-2 distribution were obtained using Independent Metropolitan algorithm.

It was discovered that two out of three regressor variables of multiple myeloma patients are significant. The covariates Bun and protein have credible regions (-0.0082, -0.0013) and (0.0282, 0.4889) respectively, which does not include zero and hence they are appropriate regressor variable for modelling survival data. The credible region for age variable is (-0.0300, 0.0494) which includes zero in it, hence is not a significant regressor variable for modelling modeling multiple myeloma data.

References

Collet, D. (1994). Modelling Survival Data in Medical Research. Chapman & Hall, London.

Collet, D. (2003). Modelling Survival Data in Medical Research, second edition. Chapman & Hall, London.

Gelman, A. and Hill, J. (2007). Data Analysis Using Regression and Multileve/Hierarchical Models, Cambridge University Press, New York.

Krall, J. M., Uthoff, V. A., and Harley, J. B. (1975). A set-up procedure for selecting variables associated with survival. Biometrics **31**, 49-57.

Okorie, I. E., Akpanta, A. C., & Ohakwe, J. (2016). The Exponentiated Gumbel Type- 2 Distribution: Properties and Application. International Journal of Mathematics and Mathematical Sciences, 2016.

Tableman, M., and Kim, J. S. (2004). Survival Analysis using S: Analysis of Time-to-Event Data. Chapman & Hall/CRC, Boca Raton.