J. Ris. & Ap. Mat. Vol. 03 No. 02 (2019) pp. 64-70
Jurnal Riset dan Aplikasi Matematika
e-ISSN: 2581-0154
URL: journal.unesa.ac.id/index.php/jram

BAYESIAN ESTIMATION OF PARETO SURVIVAL MODEL WITH INFORMATIVE PRIOR ON CENSORED DATA

SETYO WIRA RIZKI[1*], SHANTIKA MARTHA[2].

[1,2] Mathematics Department Faculty of Mathematics and Natural Sciences, Universitas Tanjungpura

Email: setyo.wirarizki@math.untan.ac.id

ABSTRACT

This research conducts a case of the cancer patients in censored data using Bayesian methodology. There are three types of loss function in Bayesian estimation method such as squared error loss function (self), linear exponential loss function (lelf) and general entropy loss function (gelf). Pareto survival model is selected as presentation data. To construct a posterior distribution, framing a likelihood function of Pareto and a prior, requires the prior distribution. An exponential distribution is chosen as a prior that describes parameter character of the Pareto. The posterior distribution is used to discover estimators in three loss functions of Bayesian methods. There are estimators held down by Bayesian self $\hat{\theta}_{BS}$, Bayesian lelf $\hat{\theta}_{BL}$ and Bayesian gelf $\hat{\theta}_{BG}$ which substance 3.79, 3.78 and 3.90 correspondingly. After getting those estimators, the hazard functions $\hat{h}_{BS}$, $\hat{h}_{BL}$ and $\hat{h}_{BG}$ and survival functions $\hat{s}_{BS}$, $\hat{s}_{BL}$ and $\hat{s}_{BG}$ can be determined. The result shows that all of survival values under Bayesian approaches are lower than the real survival value. It means the result is more trusted because as a prior, the parameter is defined more precisely than before. The hazard function confirmations a same shape in all approaches. The rates of hazard are decreasing along with survival values which show the same behavior. The curves are strictly dropping after first data. This occurrence because due to a heavy-tailed character of Pareto. The result indicates that MSE of parameter estimation under the Bayesian self, lelf and gelf are 1.3x10^{-2}, 1.2x10^{-2} and 0 respectively. The mse of survival estimation under the Bayesian self, lelf and gelf are 10^{-4}, 1.1x10^{-4} and 3x10^{-5} individually. It concludes that the Bayesian gelf is the best approximation.

Keywords: survival model, Bayesian, Pareto, prior, Exponential, heavy-tailed.

1 Introduction

There are some diseases round our life. One of them is cancer which is hazardous and making a high death risk. This disease has been suffering people in all ages. In the last years, cancer issues are growing wider in the world. In 2012, lung cancer is the most kind of cancer which attach men in Indonesia. Science, especially in Statistics and Actuarial Mathematics, provides some influences to perceive the probability of patient life which is attacked the lung cancer. Survival model is derived from both Statistics and Actuarial Mathematics. There are several estimation methods in statistics. Bayesian is one of them which usage likelihood function and prior distribution to construct posterior distribution. It is used to estimate parameter of Pareto survival model. It constructs a posterior distribution by formulating a likelihood function of Pareto and a prior. Exponential distribution is determined as a prior. A censoring is a feature that frequent in lifetime and reliability data analysis. It happens when exact lifetimes or run-outs can only be collected for a portion of the inspection units [1]. The

2010 *Mathematics Subject Classification*: 62C10

Tanggal Masuk: 06-03-19; direvisi: 01-08-19; diterima: 16-10-19

data is Pareto distributed which will be composed with an Exponential distribution as its prior to build a posterior distribution. It offers the relative weights to each parameter value after analizing the data [2]. The Bayesian inference has some approximations such as generalized non-informative prior, linear exponential loss function, Lindley approximation, general entropy loss function and squared error loss function.

2 Literature Review

2.1 Pareto Distribution

Pareto distribution is a random variable with a heavy tail (modern actuarial risk theory). There are four main properties of Pareto distribution in Survival analysis. First, the probability density function $f(t)$ presents a failure time random variable. Second, cumulative density function $F(t)$ provides the probability death less than and equal to time t. Third, survival function $s(t)$ which describes the probability of failure time is greater than the value of t. Fourth, hazard function $h(t)$ is the conditional rapid of failure quantity at time t which given survival time to t [3]. The $f(t)$, $F(t)$, $s(t)$ and $h(t)$of random variable T are formulated in equations (1), (2), (3) and (4).

$$f(t) = \frac{\alpha\theta^{\alpha}}{t^{\alpha+1}} \; ; \alpha\,,\theta > 0\,; t > \theta \tag{1}$$

$$F(t) = 1 - \left(\frac{\theta}{t}\right)^{\alpha} \tag{2}$$

$$s(t) = 1 - F(t) = \left(\frac{\theta}{t}\right)^{\alpha} \tag{3}$$

$$h(t) = \frac{f(t)}{s(t)} = \frac{\alpha}{t} \tag{4}$$

According to [4], formula of parameter estimation under squared error loss function, linear exponential loss function and general entropy loss function are defined in equation (5), (6) and (7) as next,

$$\hat{\theta}_{BS} = E(\theta) \tag{5}$$

$$\hat{\theta}_{BL} = -\frac{1}{c}\ln\left[E\left(e^{-c\theta}\right)\right] \tag{6}$$

$$\hat{\theta}_{BG} = \left[E_{\theta}(\theta)^{-k}\right]^{-\frac{1}{k}} \tag{7}$$

2.2 Bayesian Method

A way to treat an uncomplete data is censoring. It happens because some events like loss, death, or out from observation. According to [5], variables $T_1 \ldots . T_n$ represent n individual lifetimes. A time t_i is the lifetime or a censoring time. The variable $\delta_i = 0$ if $T_i > t_i$ and 1 if $T_i = t_i$ is called the censoring or status indicator for t_1. Value t_1 is obtained from $\min(T_i, C_i)\,, i = 1,2,3, \ldots, n$ where T_i is the duration of their remission measured from time of entry to study and C_i is the time between their date of entry and the end of study. The joint densiy function and joint survival function of n random variables $T_1 \ldots . T_n$ with parameter θ

is offered by $f(t_i;\theta)$ and $s(t_i;\theta)$, correspondingly. On censored data, the likelihood function for observation (t_i,δ_i) $i = 1,2,..,n$ can be expressed as,

$$\begin{aligned} L(t_i;\theta,\delta) &= \prod_{i=1}^{n}[f(t_i;\theta)]^{\delta_i}[s(t_i;\theta)]^{1-\delta_i} \\ &= \prod_{i=1}^{n}\left[\frac{\alpha\theta^{\alpha}}{t_i{}^{\alpha+1}}\right]^{\delta_i}\left[\left(\frac{\theta}{t_i}\right)^{\alpha}\right]^{1-\delta_i} \\ &= \frac{\alpha^{\sum_{i=1}^{n}\delta_i}\theta^{n\alpha}}{\prod_{i=1}^{n} t_i{}^{\delta_1+\alpha}} \end{aligned} \tag{8}$$

An exponential distribution with parameter μ is chosen as prior distribution to Pareto distribution. Let θ is a continuous random variable of exponential distribution with a parameterμ, we can present the probability density function as following,

$$f(\theta) = \mu e^{-\mu\theta} \tag{9}$$

Posterior distribution is constructed by composing of likelihood function and prior function. The posterior distribution of Pareto and exponential prior is

$$\begin{aligned} f(\theta \mid t_i) &= \frac{L(t_i;\theta,\delta)f(\theta)}{\int_0^{\infty} L(t_i;\theta,\delta)f(\theta)d\theta} \\ &= \frac{\frac{\alpha^{\sum_{i=1}^{n}\delta_i}\theta^{n\alpha}}{\prod_{i=1}^{n} t_i{}^{\delta_1+\alpha}}\mu e^{-\mu\theta}}{\int_0^{\infty}\frac{\alpha^{\sum_{i=1}^{n}\delta_i}\theta^{n\alpha}}{\prod_{i=1}^{n} t_i{}^{\delta_1+\alpha}}\mu e^{-\mu\theta}d\theta} \\ &= \frac{\theta^{n\alpha}e^{-\mu\theta}}{\int_0^{\infty}\theta^{n\alpha}e^{-\mu\theta}d\theta} \\ &= \frac{\mu^{n\alpha+1}\theta^{n\alpha}e^{-\mu\theta}}{\Gamma(n\alpha+1)} \end{aligned} \tag{10}$$

This research practices three Bayesian approaches. Firstly, the estimator of Bayes, $\boldsymbol{\hat{\theta}_{BS}}$, of $\boldsymbol{\theta}$ under the squared error loss function is the conditional mean of $\boldsymbol{\theta}$ comparative to the probability density function of posterior distribution $\boldsymbol{f(\theta \mid t_i)}$ is

$$\begin{aligned} \hat{\theta}_{BS} = E(\theta) &= \int_0^{\infty}\theta f(\theta \mid t_i)d\theta \\ &= \int_0^{\infty}\theta\frac{\mu^{n\alpha+1}\theta^{n\alpha}e^{-\mu\theta}}{\Gamma(n\alpha+1)}d\theta \\ &= \frac{\mu^{n\alpha+1}}{\Gamma(n\alpha+1)}\int_0^{\infty}\theta^{n\alpha+1}e^{-\mu\theta}d\theta \\ &= \frac{\mu^{n\alpha+1}}{\mu\mu^{n\alpha+1}\Gamma(n\alpha+1)}\int_0^{\infty}u^{n\alpha+1}e^{-u}du \end{aligned}$$

$$= \frac{\mu^{n\alpha+1}\Gamma(n\alpha+2)}{\mu\mu^{n\alpha+1}\Gamma(n\alpha+1)}$$
$$= \frac{(n\alpha+1)}{\mu} \tag{11}$$

Later, The estimator $\hat{\theta}_{BS}$ is used to find the estimator of survival function and hazard function, $\hat{s}(t)_{BS}$ and $\hat{h}(t)_{BS}$**.** They are formulated as equation (12) and (13) below,

$$\hat{s}(t)_{BS} = \left(\frac{\hat{\theta}_{BS}}{t}\right)^{\alpha}$$
$$= \left(\frac{(n\alpha+1)}{\mu t}\right)^{\alpha} \tag{12}$$

$$\hat{h}(t)_{BS} = \frac{\alpha\hat{\theta}^{\alpha}}{t^{\alpha+1}}\frac{t^{\alpha}}{\hat{\theta}^{\alpha}} = \frac{\alpha}{t} \tag{13}$$

Secondly, the estimator of Bayes, $\hat{\theta}_{BL}$, of $\boldsymbol{\theta}$under the linear exponential loss function is framed by equation (14). Moreover, the estimator $\hat{\theta}_{BL}$ is used to find the estimator of survival function and hazard function,$\hat{s}(t)_{BL}$ and $\hat{h}(t)_{BL}$**.** They are expressed as equation (15) and (16) as following,

$$\hat{\theta}_{BL} = -\frac{1}{c}\ln\left[E\left(e^{-c\theta}\right)\right]$$
$$= -\frac{n\alpha+1}{c}\ln\left(\frac{\mu}{c+\mu}\right) \tag{14}$$

$$\hat{s}(t)_{BL} = \left(-\frac{n\alpha+1}{ct}\ln\left(\frac{\mu}{c+\mu}\right)\right)^{\alpha} \tag{15}$$

$$\hat{h}(t)_{BL} = \frac{\alpha\hat{\theta}^{\alpha}}{t^{\alpha+1}}\frac{t^{\alpha}}{\hat{\theta}^{\alpha}} = \frac{\alpha}{t} \tag{16}$$

Lastly, the estimator of Bayes, $\hat{\theta}_{BG}$,of θ under the general entropy loss function is framed by equation (17). Furthermore, the estimator $\hat{\theta}_{BG}$ is used to find the estimator of survival function and hazard function, $\hat{s}(t)_{BG}$ and $\hat{h}(t)_{BG}$.They are conveyed as equation (18) and (19) as following,

$$\hat{\theta}_{BG} = \left[E_{\theta}(\theta)^{-k}\right]^{-\frac{1}{k}}$$
$$= \left[\left(\frac{\mu}{\mu+k}\right)^{n\alpha+1}\right]^{-\frac{1}{k}} \tag{17}$$

$$\hat{s}(t)_{BG} = \frac{\left[\left(\frac{\mu}{\mu+k}\right)^{n\alpha+1}\right]^{-\frac{\alpha}{k}}}{t^{\alpha}} \tag{18}$$

$$\hat{h}(t)_{BG} = \frac{\alpha\hat{\theta}^{\alpha}}{t^{\alpha+1}}\frac{t^{\alpha}}{\hat{\theta}^{\alpha}} = \frac{\alpha}{t} \tag{19}$$

An estimator which is a little biased, but having a highly focused to the parameter of interest may be desirable to an unbiased estimator that is less focused. Thus, it is desirable to have more general that allow for both biased and unbiased estimator to be compared, [6]. An bias estimator of $\hat{\theta}$ would give the bias which given by

$$b(\hat{\theta}) = E(\hat{\theta}) - \theta$$

and the mean squared error (MSE) of $\hat{\theta}$ is given by

$$MSE(\hat{\theta}) = E[\hat{\theta} - \theta]^2 \tag{20}$$

3 Result

The formulas are applied on lung cancer patient data which is taken from R versi 3.3.0. The data presents the length of time a patient has cancer in day and a censoring status. Kolmogorov-Smirnov test provides information that the data is Pareto distributed. The data runs mean value$E(t) = 121.63$. Choosing an initial value $\alpha = 2$ gives $\theta = 3.90$ by equation (21) as below

$$\begin{aligned} E(t) &= \alpha\theta^{\alpha} \int_0^{\infty} \frac{t}{t^{\alpha+1}} dt \\ &= \frac{\alpha\theta}{\alpha - 1} \end{aligned} \tag{21}$$

By calculating data of lung cancer patients, the result of parameter estimations under Bayesian self, lelf and gelf from formulas (11), (14) and (17) are presented in table 1,

Table 1.Value of parameter estimation under Bayesian self, lelf and gelf

θ	$\hat{\theta}_{BS}$	$\hat{\theta}_{BL}$	$\hat{\theta}_{BG}$
3.90	3.79	3.78	3.90

From equation (20), we can extent the MSE of each estimators of parameter beneath the three Bayseian approaches. MSE values of estimator under Bayesian self, lelf and gelf approaches can be calculated as follow,

$$MSE(\hat{\theta}_{BS}) = E[\hat{\theta}_{BS} - \theta]^2 = E[3.79 - 3.90]^2 = 1.3\text{x}10^{-2} \tag{22}$$

$$MSE(\hat{\theta}_{BL}) = E[\hat{\theta}_{BL} - \theta]^2 = E[3.78 - 3.90]^2 = 1.2\text{x}10^{-2} \tag{23}$$

$$MSE(\hat{\theta}_{BG}) = E[\hat{\theta}_{BG} - \theta]^2 = E[3.90 - 3.90]^2 = 0.00 \tag{24}$$

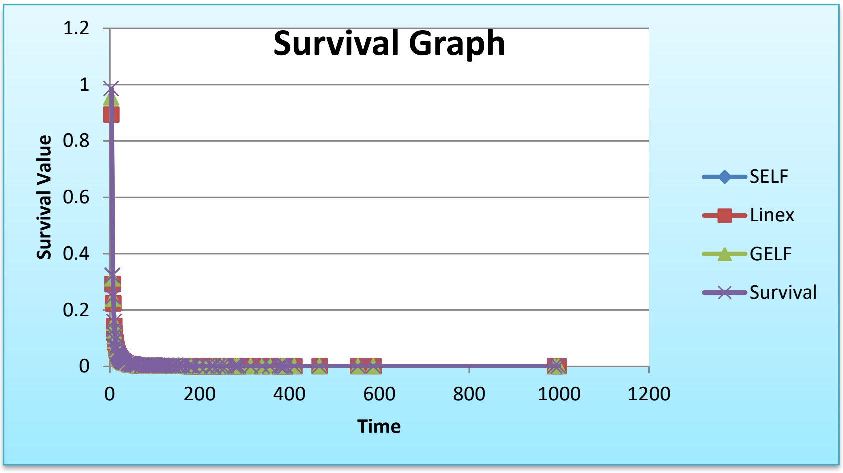
From equations (22) and (23), the results show that estimator $\hat{\theta}_{BS}$ and $\hat{\theta}_{BL}$ are biased estimators. In other hand, equation (24) indicates that $\hat{\theta}_{BG}$is an unbiased estimator. By considering the MSE, it concludes that Bayesian gelf approach is better than both of Bayesian self and gelf. From table 1, we can use the estimator values to find survival estimator values under Bayesian self, lelf and gelf. The outcome of survival estimations under Bayesian self, lelf and gelf are existed in table 2, Table 2.Value of survival estimation under Bayesian self, lelf and gelf

t	$s(t)$	$\hat{s}(t)_{BS}$	$\hat{s}(t)_{BL}$	$\hat{s}(t)_{BG}$
4	0.98	0.90	0.89	0.98
7	0.32	0.29	0.29	0.32
8	0.25	0.22	0.22	0.25
10	0.16	0.14	0.14	0.16
11	0.13	0.12	0.12	0.13
12	0.11	0.10	0.10	0.11
13	0.09	0.08	0.08	0.09
15	0.07	0.06	0.06	0.07
16	0.06	0.05	0.05	0.06
18	0.05	0.04	0.04	0.05
19	0.04	0.04	0.04	0.04
20	0.04	0.03	0.03	0.04
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
999	1.58x10^{-5}	1.44x10^{-5}	1.43x10^{-5}	1.58x10^{-5}

By referring to equation (20), the MSE of each estimators of survival underneath Bayseian approches. Calculation of the MSE values of survival estimator under Bayesian self, lelf and gelf approaces can be counted as follow,

$$MSE(\hat{s}(t)_{BS}) = E[\hat{s}(t)_{BS} - s(t)]^2 = 1.0\text{x}10^{-4} \tag{25}$$
$$MSE(\hat{s}(t)_{BL}) = E[\hat{s}(t)_{BL} - s(t)]^2 = 1.1\text{x}10^{-4} \tag{26}$$
$$MSE(\hat{s}(t)_{BG}) = E[\hat{s}(t)_{BG} - s(t)]^2 = 0.00 \tag{27}$$

The result of survival estimation can be described on a graphic like as graph 1 following,



Graph 1 Survival probability estimation graph under Bayesian approaches

The table 2 and graph 1 give us information about different results of survival chance of lung cancer patients under Bayesian approaches. The three survival estimator curves coincide with the survival value curve. This is because the biased level of estimators are very small and there are even unbiased estimator. The graph shows there is an extremely decreasing result after first data then getting smoother after second data. It happens because the Pareto distribution is a heavy-tailed distribution.

4 Conclusion

In brief, the Bayesian gelf gives an unbiased estimator of parameter of survival model. It is disclosed by giving a same value of both of parameter and estimator. The results of estimating survival values using the Bayesian self and linex approximation tends to being smaller than the real survival value and Bayesian gelf estimator . It is caused by influence of prior and the type of approach. By means of considering the MSE values, we can conclude that Bayesian gelf approach is better than the other two methods.

REFERENCES

[1] Guure, C. B., & Ibrahim, N. A. (2012). Bayesian Analysis of the Survival Function and Failure Rate of Weibull Distribution with Censored Data. *Mathematical Problems in Engineering*.

[2] Bolstad, W. M. (2007). *Introduction to Bayesian Statistics* (Second ed.). New Jersey: John Wiley & Sons, Inc

[3] London, D. (1988). *Survival Models and their Estimation.* Winsted: Actex Publicatins, Inc.

[4] Guure, C. B., & Ibrahim, N. A. (2014). Approximate Bayesian Estimates of Weibull Parameters with Lindley's Method. *Sains Malaysiana*, *43*(9), 1433-1437.

[5] Metiri, F., Zeghdoudi, H., & Remita, M. R. (2016). On Bayes Estimatesof Lindley Disitribution under Linex Loss Function: Informative and Non Informative Priors. *Global Journal of Pure and Applied Mathematics*, *12*(1), 391-400.

[6] Bain, L. J., & Engelhardt, M. (1991). *Introduction to Probability and Mathematical Statistics.* California: Duxbury Press.