Regularization Bayesian Lasso Right Censored Regression Model

Zainab Saadoon Oleiwi
Universirty of Al-Qadisiyah
admin.stat21.16@qu.edu.iq

Bahr kadhim mohammed
Universirty of Al-Qadisiyah
Bahr.mahemmed@qu.edu.iq

Abstract

This paper focuses on studying the method of regularization called the Lasso method from the Bayesian theory point of view. Three models have been employed to represent the prior distribution of the regression parameter (Laplace distribution), the first model assumed the use of a scale mixture of normal distribution mixing and the exponential distribution. The proposed model is the second representation of the scale mixture of uniform distribution mixing with the standard exponential distribution, and the third scale mixture of uniform distribution mixing with the gamma distribution. The three models have been applied in the right-censored regression. The Bayesian estimation has been conducted by implementing the Gibbs sampling algorithm with three simulation examples via the R programming language with different sample sizes and different variance values for errors. In order to demonstrate the efficiency of the proposed method, this method was employed on real data, the data set is a sample with the right-censored response variable that represents the level of urea in the blood with a set of explanatory variables. The results showed that the proposed method comparable with other methods in terms of prediction accuracy and variables selection procedure.

Keywords: Bayesian Lasso, right censored data, variable selection, Gibbs sampler

1. Introduction

Regression analysis is the most important branch of statistics which is concerned with building the mathematical relationship between the dependent variable and the independent variables, this relationship is represented as a linear formula called the regression equation, as its accuracy depends on the correctness of estimating its parameters. The general form of a linear regression model is:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

where $\boldsymbol{y}$ is a vector of $(n \times 1)$ of responses, $\boldsymbol{X}$ is a matrix with dimension $(n \times p)$ of predictors, $\boldsymbol{\beta}$ is a vector $(p \times 1)$ of unknown parameters, and $\boldsymbol{\epsilon}$ is a vector $(n \times 1)$ of random errors.

The OLS method is one of the most important and common methods for estimating the parameters of the linear regression model. This method is characterized by good characteristics that made it one of the best and most widely used methods. This method is based on the principle of minimizing the sum of the squares of errors to the least possible (Balestra, 1970). The mathematical formula for obtaining the (OLS) estimator for the parameters of the regression model using matrices is as follows:

$$\widehat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y},$$

The (OLS) method gives the best unbiased linear estimate (BLUE) with the least variance of the model parameters.

The variable selection process refers to a collection of tasks in which the goal is to find the optimal subset of relevant variables that can be utilized to make precise modifications to the outcomes of a given dependent variable. Identifying essential

and influential factors on the dependent variable might be challenging when the number of variables is too great. As a result, in the data analysis, the (VS) feature was deemed important. To overcome the disadvantages of the (OLS) method and other classical methods like the forward method, backward method, stepwise method, and all subset regression are suitable in cases where the number of independent variables (p) is large. Therefore, regularization methods have been proposed such as the Lasso method and others to obtain the best estimate for the unknown parameter from between all possible estimates. The Lasso a method proposed by researcher Tibshirani in 1996 works on selecting the variables and estimating parameters of the regression model at the same time. It's a method (OLS) and but, it a restricted. In contrast to (OLS), the Lasso estimation method is biased but more accurate according to the following formula:

$$\widehat{\boldsymbol{\beta}}_{LASSO} = \underbrace{arg\ min}_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \lambda \sum_{j=1}^{p} \left|\boldsymbol{\beta_j}\right|.$$

Because of the normal of this constraint, the LASSO method reduces some estimated parameters and makes other parameters equal to zero, thus reducing the variance of errors. And it becomes easy to interpret the regression model (Tibshirani, 1996; Savin, 2013).

The Bayesian analysis became more popular because of the development of computer approximations to integrals and the appearance of easy-to-use programs to implement these arithmetic operations. And the use of Bayesian statistics was not limited to the development of wide research in Bayesian methodology, but also in the use of Bayesian methods to process many problems in applied domains. (Rencher & Schaalje, 2008; Chand, 2012)

The limited dependent variables in regression models mean that there is a limit to the dependent variable. And some independent variables reach that limit. Where the dependent variable is observed within a specified range, while the independent variables are observed within an open range. Limited dependent variable models address two issues important censored and truncation. A limited dependent variable y_i^* is a continuous variable with a lot of independent variables repeated at the lower or upper bound (Tobin, 1958; Maddala, 1987).

In this paper, we introduced a mixed representation of the Laplace distribution by performing transformations and mathematical operations, which was obtained through the uniform continuous distribution $(\frac{-\sigma^2}{\lambda}, \frac{\sigma^2}{\lambda})$ multiplied by the standard exponential distribution (z). It was employed for Bayesian Lasso regression for left and right censored data. The results show that the proposed method performs very well compared with the classical methods for left and right-censored data. And we have proposed a Bayesian regularization method for left and right-censored responses based on the Bayesian regularized method of Park & Casella (2008). Also, we have proposed a Bayesian regularization method for left and right-censored responses based on the Bayesian regularized method of Mallick & Yi (2014). In practice, the results show that the proposed methods perform very well in terms of convergence.

2. Right Censored Data

A data point is upper than a particular value but is an unknown. If the latent variable y_i^* is lower than the limit and the limit for the censored observations, the real value for the dependent variable y_i is observed. If the dependent variable's actual values are less than the upper limit, they are observed (Koul et al., 1981;

Kohler et al., 2002). The structural formulation of the right censored linear regression is defined by:

$$y_i = \begin{cases} y_i^* & if \quad y_i^* < y_U, \\ y_U & if \quad y_i^* \geq y_U, \end{cases} \tag{1}$$

or equivalently,

$$y_i = min\{y_i^*, y_U\},$$

where $$y_i^* = \boldsymbol{x_i'\beta} + \varepsilon_i,$$

y_i^* is the latent variable or the unobservable variable. Furthermore, the following write down represents the Gibbs sample algorithms:

3. The Proposed Scale Mixture

Based on the following mathematically fact,

$$\int_{w > \frac{|x|}{\sigma^2}} \lambda e^{-\lambda w}\, dw = e^{-\frac{\lambda|x|}{\sigma^2}} \tag{2}$$

we can propose the following scale mixture formula. In (2), let $x = \beta,\ \lambda w = z$, and by multiply both sides by $\frac{\lambda}{2\sigma^2}$, we get

$$\frac{\lambda}{2\sigma^2} \int_{\frac{z}{\lambda} > \frac{|\beta|}{\sigma^2}} \lambda e^{-z} \frac{1}{\lambda} dz = \frac{\lambda}{2\sigma^2} e^{-\frac{\lambda|\beta|}{\sigma^2}}$$

$$\frac{\lambda}{2\sigma^2} e^{-\frac{\lambda|\beta|}{\sigma^2}} = \int_{z > \frac{\lambda|\beta|}{\sigma^2}} \frac{\lambda}{2\sigma^2} e^{-z}\, dz \tag{3}$$

so, the formulation (3) is the scale mixture of standard exponential mixing with uniform $(\frac{-\sigma^2}{\lambda}, \frac{\sigma^2}{\lambda})$.

3.1 The Hierarchical Prior Model of Right-Censored Data

Based on the proposed scale mixture (3), and (1). The hierarchical prior model is formulated as follows:

$$y_i = \begin{cases} y_i^* & if \;\; y_i^* < y_U, \\ y_U & if \;\; y_i^* \geq y_U, \end{cases}$$

$$y_i^* | \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n),$$

$$\boldsymbol{\beta} | \; \sigma^2, \lambda \sim Uniform\left(-\frac{\sigma^2}{\lambda}, \frac{\sigma^2}{\lambda}\right), \tag{4}$$

$$\sigma^2 \sim \pi(\sigma^2) \; d\sigma^2,$$

$$\lambda \sim Gamma(c \, , d),$$

$$z \sim Exp(1).$$

Where $\boldsymbol{X}$ is the standardized covariate matrix, and $\boldsymbol{y}^*$ are the centered unobserved response variable values.

3.2 The Gibbs Sampling Algorithms

Sampling parameters of the right censored regression model (1), unobserved Variables, the hierarchical prior model (4) guide us to the exact Gibbs sampler with the following steps:

1. Sampling $\boldsymbol{y}^*$: we draw samples from

$$y_i^* | y_i \, , \boldsymbol{\beta} \sim \begin{cases} y_i & if \;\; y_i^* < y_U, \\ N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n) & if \;\; y_i^* \geq y_U, \end{cases}$$

2. Sampling $\boldsymbol{\beta}$: we draw samples from

$$\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},z,\lambda,\sigma^2 \sim N_k(\widehat{\boldsymbol{\beta}}_{OLS},\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})\prod_{j=1}^{k} I\left\{\frac{-z_j\sigma^2}{\lambda} < \beta_j < \frac{z_j\sigma^2}{\lambda}\right\}$$

3. Sampling σ^2: we draw samples from

$$Inverse - Gamma\left(\frac{n}{2}+q+k, \frac{(\boldsymbol{y}^*-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}^*-\boldsymbol{X}\boldsymbol{\beta})}{2+\theta}\right)$$

4. Sampling z: we draw samples from

$$\prod_{j=1}^{k} standard\ exponential\ I\left\{z_j > \frac{\lambda|\beta_j|}{\sigma^2}\right\}$$

5. Sampling λ: we draw samples from

$$Gamma\ (k+c,d)\prod_{j=1}^{k} I\left\{\lambda < \frac{z_j\sigma^2}{|\beta_j|}\right\}$$

4. Simulation Study and Real Data

4.1 Simulation Study

In this section, we demonstrate the prediction accuracy of the methods: linear right-censored regression (RCR), Bayesian LASSO right-censored regression (BLRCR), the proposed Bayesian LASSO right-censored regression (NBLRCR), and Bayesian LASSO right-censored regression using scale mixture uniform (BLRCRsmu). The outcome variable is centered and the covariates are standardized to have 0 means and unit variances before applying the above methods. For the prediction accuracy, we evaluate the median of mean squared errors (MMSE) for the simulated studies based on 100 replications.

Example 1 (Right-censored with sparse case)

In this example, except that we set $\beta^{10\times1} = (6,1,0,0,3,0,0,0,0,0)'$, $\sigma^2 = \{1,1.5,2\}$ and we generate data from the correct model

$$y_i = min(5, y_i^*), \qquad i = 1, \dots, n,$$

$$y_i^* = \boldsymbol{x_i'\beta} + \varepsilon_i, \qquad i = 1, \dots, n,$$

The results are listed in Table (1). The results show that the proposed Bayesian LASSO right-censored regression (NBLRCR) performs very well compared to other methods in the comparison. It has the smallest MMSE in 7 out of 9 experimental results. The Bayesian LASSO right-censored regression (BLRCR) also performs well compared to other methods in the comparison. It has the smallest MMSE in 2 out of 9 experimental results.

Table 1: Median mean squared error (MMSE) and their associated standard deviations (SD) are listed in the parentheses for Example (1). All results are averaged over 100 replications.

(n_t, n_p, σ^2)	RCR	BLRCR	PBLRCR	BLRCRsmu
(100,200,1)	0.1153(0.0336)	0.1409(0.1124)	**0.0878**(0.0410)	0.1047(0.0333)
(150,200,1)	0.2331(0.1111)	0.2961(0.1212)	**0.1320**(0.0315)	0.1981(0.0838)
(200,200,1)	0.4286(0.2022)	0.5258(0.3168)	**0.2784**(0.1684)	0.3683(0.1904)
(100,200,1.5)	0.0651(0.0134)	0.0617(0.0329)	**0.0377**(0.0078)	0.0565(0.0120)
(150,200,1.5)	0.1238(0.1136)	0.1432(0.0435)	**0.0737**(0.0415)	0.1078(0.0908)
(200,200,1.5)	0.3585(0.2697)	0.3597(0.0740)	**0.2337**(0.1508)	0.3010(0.2128)
(100,200,2)	0.0560(0.0174)	0.0724(0.0301)	**0.0520**(0.0157)	0.0537 0.0174)
(150,200,2)	0.1886(0.0796)	**0.1197**(0.0328)	0.1266(0.0617)	0.1691(0.0749)
(200,200,2)	0.1889(0.0801)	**0.1201**(0.0331)	0.1288(0.0612)	0.1696(0.0751)

Convergence of the corresponding our Gibbs sampler methods was assessed by trace plots of the simulated draws. The trace plots Figures (1 – 3) shows that our methods converge very fast.

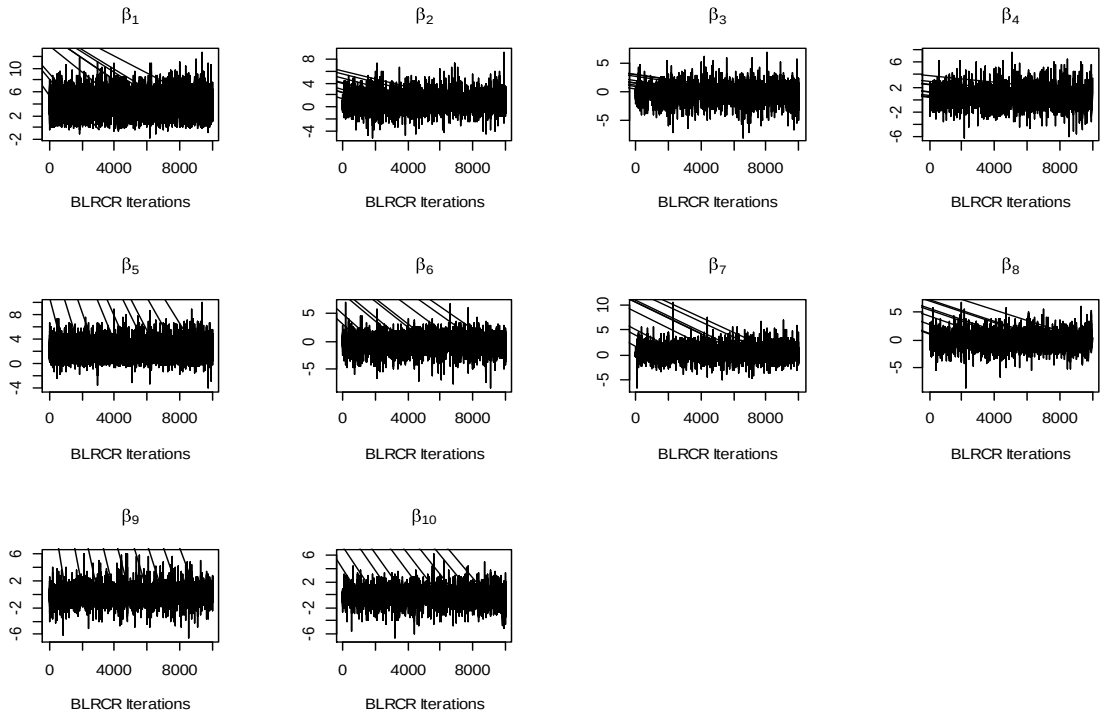

Figure 1: Trace plots of parameters in simulation 1 using BLRCR method.

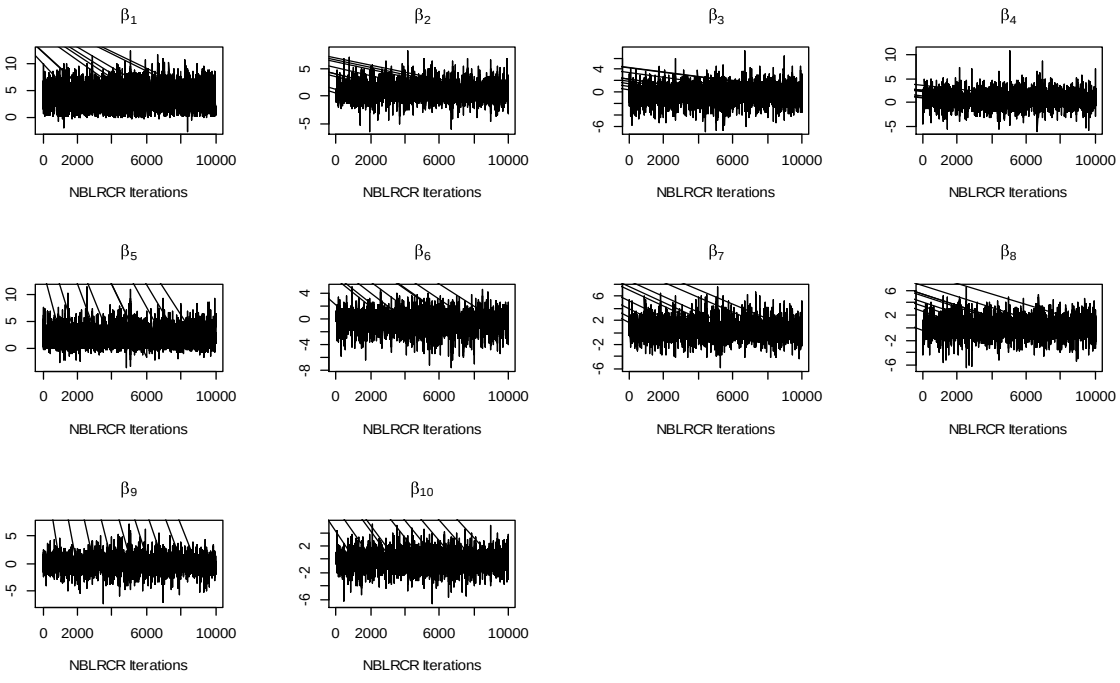

Figure 2: Trace plots of parameters in simulation 1 using PBLRCR method.

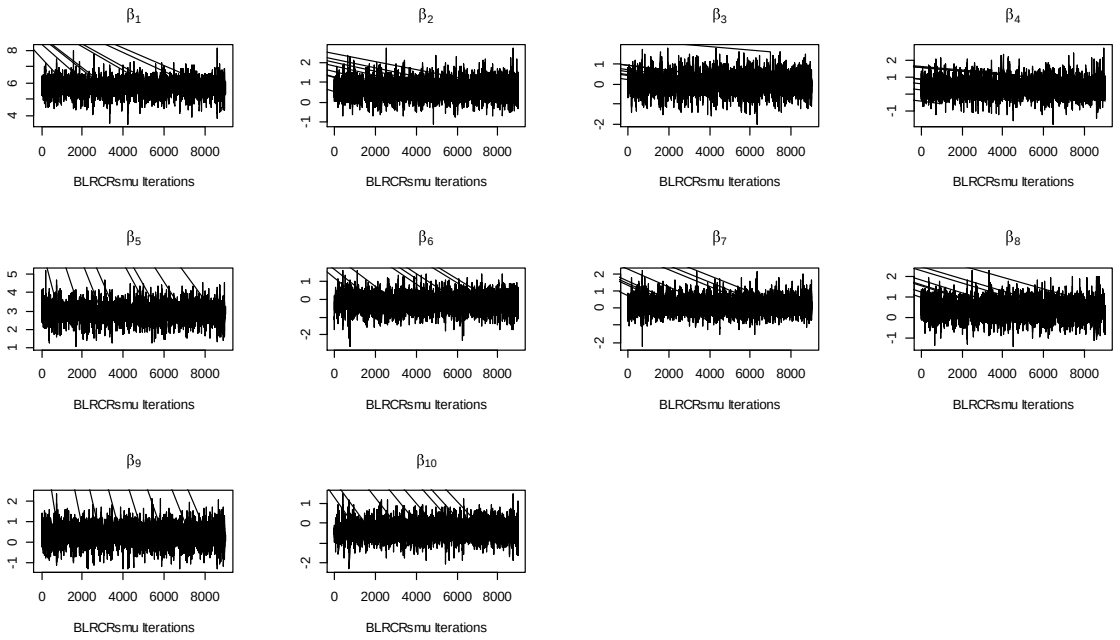

Figure 3: Trace plots of parameters in simulation 1 using BLRCRsmu method.

Example 2 (Right-censored with dense case)

This example is similar to Example (1) except that we set $\beta^{10\times1} = (6,1,1,1,1,1,1,1,1,1)'$, $\sigma^2 = \{1,1.5,2\}$ and we generate data from the correct model

$$y_i = min(5, y_i^*), \qquad i = 1, \dots, n,$$

$$y_i^* = \boldsymbol{x_i'\beta} + \varepsilon_i, \qquad i = 1, \dots, n,$$

The results are listed in Table (2). The results show that the proposed Bayesian LASSO right-censored regression (NLRCR) performs very well compared to other methods in the comparison. It has the smallest MMSE in 7 out of 9 experimental results. The Bayesian LASSO right-censored regression using scale mixture uniform (BLRCRsmu) also performs well compared to other methods in the comparison. It has the smallest MMSE in 2 out of 9 experimental results.

Table 2: Median mean squared error (MMSE) and their associated standard deviations (SD) are listed in the parentheses for Example (2). All results are averaged over 100 replications.

(n_t, n_p, σ^2)	RCR	BLRCR	PBLRCR	BLRCRsmu
(100,200,1)	0.2087 (0.0372)	0.5022(0.1124)	**0.1485** (0.0490)	0.1506(0.0376)
(100,200,1)	0.3448 (0.1619)	0.7973(0.1993)	**0.3309** (0.1592)	0.3348(0.1560)
(100,200,1)	0.6431 (0.2490)	1.4096(0.6720)	**0.6179** (0.3090)	0.6233(0.2321)
(150,200,1.5)	0.0832 (0.0436)	0.1600(0.0620)	0.0890 (0.0120)	**0.0813**(0.0400)
(150,200,1.5)	0.2122 (0.0480)	0.3839(0.0727)	**0.1677** (0.0262)	0.1708(0.0483)
(150,200,1.5)	0.3085 (0.0758)	0.7735(0.2405)	**0.2688** (0.0827)	0.2758(0.0886)
(200,200,2)	0.0511 (0.0172)	0.0768(0.0156)	0.0533 (0.0070)	**0.0496**(0.0157)
(200,200,2)	0.1106 (0.0420)	0.2071(0.0742)	**0.1009** (0.0382)	0.1019(0.0418)
(200,200,2)	0.2659 (0.1149)	0.4753(0.0737)	**0.2492** (0.0718)	0.2539(0.1031)

Example 3 (Right-censored with very sparse case)

This example is similar to Example (1) except that we generate data from the correct model

$$y_i = min(5, y_i^*), \qquad i = 1, \dots, n,$$

$$y_i^* = \boldsymbol{x_i'\beta} + \varepsilon_i, \qquad i = 1, \dots, n,$$

The results are listed in Table (3). The results show that the Bayesian LASSO right-censored regression (BLRCR) performs very well compared to other methods in the comparison. It has the smallest MMSE in 6 out of 9 experimental results. The proposed Bayesian LASSO right-censored regression (NBLRCR) also performs well compared to other methods in the comparison. It has the smallest MMSE in 3 out of 9 experimental results.

Table 3: Median mean squared error (MMSE) and their associated standard deviations (SD) are listed in the parentheses for Example (3). All results are averaged over 100 replications.

(n_t, n_p, σ^2)	RCR	BLRCR	PBLRCR	BLRCRsmu
(100,200,1)	0.1075(0.0516)	**0.0444**(0.0414)	0.0608(0.0277)	0.0821(0.0427)
(100,200,1)	0.2998(0.1106)	**0.1414**(0.0549)	0.1509(0.0547)	0.2212(0.0794)
(100,200,1)	0.2587(0.1850)	1.1212(0.0852)	**0.1300**(0.0687)	0.1690(0.1198)
(150,200,1.5)	0.1240(0.0587)	0.0777(0.0471)	**0.0743**(0.0414)	0.1069(0.0549)
(150,200,1.5)	0.2465(0.1547)	**0.0971**(0.0546)	0.1496(0.0899)	0.2000(0.1225)
(150,200,1.5)	0.2389(0.1096)	**0.0954**(0.0717)	0.1461(0.0670)	0.1685(0.0858)
(200,200,2)	0.0587(0.0234)	0.0406(0.0201)	**0.0383**(0.0191)	0.0510(0.0214)
(200,200,2)	0.1211(0.0352)	**0.0438**(0.0104)	0.0836(0.0333)	0.0957(0.0317)
(200,200,2)	0.2687(0.1187)	**0.0903**(0.0476)	0.1815(0.0853)	0.2004(0.0929)

4.2 Real Data

Data for 62 patients were obtained. These data were collected from Al-Hashimiya General Hospital. The research variables consist of a dependent variable and 10 independent variables, which are;

Urea level in blood (y_i)**:** (Uremia) is caused by extreme and usually irreversible damage to your kidneys. This is usually from chronic kidney disease. The kidneys are no longer able to filter the waste from your body and send it out through your urine. Instead, that waste gets into your bloodstream, causing a potentially life-threatening condition. The normal percentage of urea in blood around 6 to 24 mg/dL (2.1 to 8.5 mmol/L) is considered. The set of independent variables that can affect the dependent variable: Age (x_1), Urinary tract obstruction (x_2), Congestive heart failure (x_3), Having a heart attack (x_4), Gastrointestinal bleeding (x_5), Drought (x_6), Severe burns (x_7), Pharmaceuticals (x_8), Sugar percentage (x_9), Blood fat levels (x_{10}).

Table 4: showed the estimation of parameters.

Descriptive variables	Variables	PBLRCR	BLRCRsmu	BLRCRsmn
Age	x_1	0.000	0.003	0.000
Urinary tract obstruction	x_2	0.654	0.793	0.543
Congestive heart failure	x_3	0.000	0.000	0.065
Having a heart attack	x_4	0.763	0.439	0.652
Gastrointestinal bleeding	x_5	0.000	0.000	0.005
Drought	x_6	0.732	0.874	0.609
Severe burns	x_7	0.000	0.000	0.070
Pharmaceuticals	x_8	2.210	3.095	0.054
Sugar percentage	x_9	0.854	0.986	0.549
Blood fat levels	x_{10}	0.000	0.000	0.005
MSE		**13.98**	16.43	17.86

We can see that the proposed model gave least the value for MSE, is 13.98 , and the from above table the estimation of parameters were taken from the subsequent distributions of the proposed model, by adding a threshold point to zeroing because Bayesian methods do not zero, and the proposed method has reduced many unimportant variables such as making a variable selection in the proposed model in the five variables (Age, Congestive heart failure, Gastrointestinal bleeding, Severe burns, and Blood fat levels) where the parameters were ($x_1 = 0,\ x_3 = 0,\ x_5 = 0,\ x_7 = 0, and\ x_{10} = 0$).

5. Conclusions

In this paper, we have proposed several Bayesian methods for variable selection and parameter estimation in linear regression models with right-censored data. Some advantages over old approaches include fast convergence Gibbs sampler, efficient Gibbs sampler computation techniques, and the use of data augmentation to allow right-censored responses. The criterion of the median of mean squares error has been used to test the performance of the different methods; the results showed that the proposed method is comparable with the other methods.

Bibliography

Balestra, P. (1970). On the efficiency of ordinary least-squares in regression models. Journal of the American Statistical Association, 65(331), 1330-1337.

Chand, S. (2012, January). On tuning parameter selection of LASSO-type methods-a monte carlo study. In Proceedings of 2012 9th international Bhurban conference on applied sciences & technology (IBCAST) (pp. 120-129). IEEE.

Kohler, M., Máthé, K., & Pintér, M. (2002). Prediction from randomly right censored data. Journal of Multivariate Analysis, 80(1), 73-100.

Koul, H., Susarla, V., & Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. The Annals of statistics, 1276-1288.

Maddala, G. S. (1987). Limited dependent variable models using panel data. Journal of Human resources, 307-338.

Mallick, H., & Yi, N. (2014). A new Bayesian lasso. Statistics and its interface, 7(4), 571-582.

Park, T., & Casella, G. (2008). The bayesian Lasso. *Journal of the American Statistical Association*, *103*(482), 681-686.

Rencher, A. C., & Schaalje, G. B. (2008). Linear models in statistics. John Wiley & Sons.

Savin, I. (2013). A comparative study of the Lasso-type and heuristic model selection methods. Jahrbücher für Nationalökonomie und Statistik, 233(4), 526-549.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, 24-36.