Shivangee Misra, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 11, Issue 2, (Series-III) February 2021, pp. 20-26

RESEARCH ARTICLE OPEN ACCESS

Robust Bayesian Model Selection for Simple Linear Regression Using Log-Normal Error Disturbances

Shivangee Misra *, Rajeev Pandey**
**(Department of Statistics, University of Lucknow, India)*
*** (Department of Statistics, University of Lucknow, India)*

ABSTRACT
The present study provides a Robust Bayesian Prediction model for the prediction of estimates for outcome variable for a given value of explanatory variable in a Simple Linear Regression Model with error term distributed log-normally. Markov chain Monte Carlo (MCMC) simulation techniques are used to obtain the posterior estimates of unknown parameters and the predictive estimates of the response variable are obtained using extension of simulation in the regression techniques under the violation of the normality assumption. The present paper provides Bayesian linear regression approach using Gibbs sampling to make prediction about the response variable for a given value of explanatory variable under the assumption of non-Gaussian error terms. As in a Bayesian paradigm, the population parameters are treated as random variables and often consists of complex statistical models, the OpenBUGS (i.e., Bayesian inference Using Gibbs Sampling) software application is brought in use for the estimation of various parameters using the R2OpenBUGS package of RStudio.

Keywords **-** Bayesian linear regression, Non-normal distribution, Posterior Predictive Estimates, Gibbs sampling.

Date of Submission: 25-01-2021 Date of Acceptance: 10-02-2021

I. INTRODUCTION

The real-life data in many fields of Health, Education, and the Social Sciences sectors mostly deviate from normal behavior and the distribution of random noise depart from the normality assumption. In such cases making predictions of the response variable for a particular explanatory variable through the Ordinary Least Squares regression might give us misleading results. Under such circumstances, the proposed Bayesian prediction model can be easily applied for providing predictive estimates of the response variable for a given value of explanatory variable. The present paper provides a development of a Bayesian prediction model in simple linear regression (SLR) for the outcome variable when the random noise follows a log normal distribution instead of normal error. The consistency of results while predicting the future outcome variable is analysed and under such circumstances, a Robust Bayesian Prediction model is developed for the prediction of the outcome variable for given values of explanatory variable. The robustness of the model is observed in terms of Deviance Information Criterion (DIC) which ensures the future predictions and do not exert an undue influence on the inferences of the proposed model. In the present paper, Gibbs sampling which is a MCMC technique is suitably used. The open-source variant of WinBUGS (i.e., Bayesian inference Using Gibbs Sampling) is employed for the Bayesian analysis of such complex statistical models using Markov chain Monte Carlo (MCMC) methods.

II. METHODOLOGY

2.1 MODEL SPECIFICATION

Let the data be a set of n observations $\{\mathbf{z_i}\}$ of the dependent variable, with i =1, . . ., n and their associated vector $\{\boldsymbol{x_i}\}$ being the predictor. We consider a linear regression model with an error term distributed according to lognormal distribution:

$$\mathbf{z_i} = \boldsymbol{\alpha} + \boldsymbol{\beta x_i} + \boldsymbol{\varepsilon_i} \; ; \varepsilon_i \sim lognormal(0, \sigma^2) \tag{2.1}$$

and three unknown parameters α, β and σ^2

z_i: i^{th} outcome variable

x_i: i^{th} explanatory variable

ε_i: Random noise associated with the i^{th} variable

α: Intercept of the model

β: Regression Coefficient of the model

2.2 Model Transformation

As suggested by Zellner **[1]**, the response variable is transformed into a normal variable by using the natural log transformation considering $\mathbf{y_i = \ln z_i}$ as the new transformed variable. Now, the transformed model is as follows:

$$\mathbf{y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta x}_i + \boldsymbol{\varepsilon}_i \; ; \quad \varepsilon_i \sim Normal(0, \sigma^2) \tag{2.2}$$

2.3 Bayesian Analysis

In Bayesian paradigm, **Θ** is the unknown quantity. Since there is uncertainty about the parameter, it is regarded as a random variable and a probability distribution is assigned to it, called the Prior distribution. The posterior density function includes the prior information about **Θ** in terms of prior distribution and the information contained in the data via likelihood function. Its distribution function is obtained by the following Bayes formula:

$$P(\boldsymbol{\Theta}|data) = \frac{P(data|\boldsymbol{\Theta}) \times P(\boldsymbol{\Theta})}{P(data)}. \tag{2.3}$$

where, P(Θ)is the prior distribution for the unknown parameter i.e., the strength of our belief about the parameters based on degree of belief. ***P(data*| Θ*)*** is the distribution function of the observations given the prior belief. ***P(data)*** is the evidence. This is the probability of data as determined by summing (or integrating) across all possible values of Θ, weighted by how strongly we believe in the values of Θ. The ***P(Θ|data)*** is the Posterior distribution representing the posterior beliefs about the parameters.

2.4 Prior Specification

The vague normal prior for the parameters $\alpha \; and \; \beta$ and an inverse-gamma distribution on the variance σ^2 were proposed by Spiegelhalter **[2]**. Thus, the prior specification for the model can be easily employed here as follows:

$$\alpha \sim N(0,1000)$$
$$\beta \sim N(0,1000)$$
$$\tau \sim gamma(0.001,0.001) \; ; \left(\tau = \frac{1}{\sigma^2}\right) \text{ with shape and rate parameters respectively.} \tag{2.4}$$

2.5 Bayesian Prediction

To estimate a future outcome value for a given pre-specified explanatory variable, the prediction in Bayesian framework for the $(n+1)^{th}$ future response variable will involve the posterior knowledge of the parameters in the model, on clubbing the likelihood and the prior. These predictions are outcome values simulated from the posterior predictive distribution, which is the distribution of the unobserved (future) data given the set of observations of explanatory variable.

Let, $\boldsymbol{y_i}*$ (for i= n+1) be the next predicted outcome variable for a given value of $\boldsymbol{x_i}*$ (for i =n+1):

$$\boldsymbol{y_i}* = \boldsymbol{\alpha} + \boldsymbol{\beta x_i}* + \boldsymbol{\varepsilon_i} \tag{2.5}$$

This $\boldsymbol{y_i}*$, the value of future predicted outcome variable for a given value of explanatory variable $\boldsymbol{x_i}*$, for the $(n+1^{th})$ observation is predicted using the simulation analysis.

III. NUMERICAL ILLUSTRATION

To demonstrate how to perform the analysis in R2Open BUGS package, the following real data set has been obtained from the Diabetes Care Centre, Lucknow.

The dataset considered consists of N=287 patients. There are 10 explanatory variables like Age, Gender, TC, TG, HDL, BMI etc. In the present study, the BMI is taken as a significant explanatory variable for predicting the Median Stiffness of Liver, the outcome variable.

Table 1: Numerical Illustration

Sr. No.	Age	Gender	TC	TG	HDL	LDL	VLDL	BMI	HBA1C	Duration	Median Stiffness
1	20	0	120.10	145.90	23.90	100.90	37.99	34.58	5.90	12.88	4.83
2	21	0	150.61	215.89	120.34	210.34	44.49	27.39	10.10	15.87	13.58
3	48	0	154.40	112.30	42.10	99.80	22.90	29.05	13.35	10.65	3.21
4	51	1	187.81	218.15	58.24	107.95	51.19	27.40	10.08	12.91	4.84
5	51	0	178.51	188.46	59.28	96.08	45.60	26.88	9.53	9.49	3.40
6	53	1	170.80	179.60	67.20	157.20	67.89	27.37	6.20	14.93	12.11
7	51	0	182.18	200.18	58.87	100.76	42.50	27.09	9.75	10.04	3.81
…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…
287	18	0	129.71	127.33	36.78	121.78	43.04	25.34	13.20	13.28	15.65

3.1 SUMMARY STATISTICS OF THE DATASET

The following output is generated in RStudio for all the variables.

Table 2: Summary statistics of the dataset

	Min	1st Quantile	Median	Mean	3rdQuantile	Max
Age	18.05	36.15	51.45	46.96	54.00	91.00
Gender	0	0	0	0.39	1	1
TC	36.00	144.80	180.00	175.30	187.80	376.40
TG	40.00	161.50	199.00	195.50	216.80	511.30
HDL	23.90	53.63	59.02	78.97	100.75	210.92
LDL	26.48	100.83	115.79	138.68	185.59	311.20
VLDL	15.68	36.94	43.34	49.68	47.64	185.49
BMI	17.68	25.98	27.14	**27.03**	28.09	39.21
HB1AC	5.00	7.50	9.55	9.27	10.02	16.40
Duration	1.00	9.68	11.38	10.99	13.30	17.00
Median Stiffness	3.21	3.64	5.37	**7.156**	0.15	16.60

3.2 PLOTTING THE DISTRIBUTION OF THE RESPONSE VARIABLE

The following figure illustrates empirical density plot of the outcome variable (before log transformation) and its cumulative distribution.

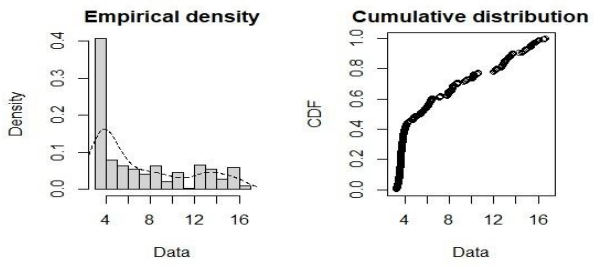

Fig. 1: Empirical and Cumulative Distribution plot of the Response variable

The plot of the response variable (Median stiffness) is highly positively skewed and resembles with a lognormal distribution. Furthermore, the betterment of fit of the response variable under the non-normality assumption for the error terms of the model (2.1) is computed. The table below illustrates goodness of fit test using Scipy library of Python.

Table 3: Distributions listed by betterment of fit

Distribution	*chi_square*
lognorm	29.651
invgauss	33.005
weibull_min	41.827
gamma	44.430
pearson3	68.484
expon	81.620
beta	138.834
triang	191.483
weibull_max	237.978
norm	361.013
uniform	363.129

From this finding, it is evident that the log normal is observed to be of best fit with the least chi square value 29.65 as compared with the normal distribution having chi square value as 361.01. Thus, it clearly indicates the model contains errors distributed log-normally.

IV. SIMULATION PROCEDURE

The RStudio, an integrated development environment (IDE) for R, a programming language for statistical computing and graphics is employed for the executions of functions interactively. As suggested by Edward Greenberg **[3]**, the Bayesian Analysis with the help of OpenBUGS (Bayesian inference Using Gibbs Sampling), a popular software for analyzing complex statistical models using MCMC methods is employed. Here, we propose R2OpenBUGS package recommended by Gelman et al. **[4]** for the robust prediction for the model (2.5) under the violation of normality assumption.

A BUGS model using the prior specification (2.4) and the regression model (2.2), the transformed model is set up in R using OpenBUGS software and the following analysis has been obtained.

4.1 SUMMARY STATISTICS OF THE BUGS MODEL

The following output table shows the posterior mean and standard deviation, a set of five quantiles for the parameters of the model, the Deviance Information Criterion (DIC) and the effective number of parameters (pD). The thinning interval is unity and three chains each of 9900 sample size are performed for 10000 iterations. For each parameter, the respective convergence diagnostics, i.e., Rhat is also computed.

Table 4: Posterior estimates and the convergence diagnostics of the parameters

	Mean	*sd*	*25%*	*50%*	*97.5%*	*Rhat*
alpha	1.810	0.032	1.789	1.810	1.874	**1.00094**
beta	0.035	0.011	0.027	0.035	0.058	**1.00092**
sigma	0.544	0.023	0.528	0.543	0.591	**1.001**
Deviance	464.806	2.465	463	464.20	471.102	**1.001**
pD=3						
DIC=467.8						

It is hereby observed that the potential scale reduction factor i.e., Rhat converges to unity indicating that the marginal behavior of the chain is sufficiently close to stationarity.

4.2 TRACE PLOTS AND DENSITY PLOTS OF POSTERIOR DISTRIBUTION

4.2.1 The following output is generated within the CODA **[5]**, a designed package for R to take BUGS output as input.

Number of chains= 3 (green, red, blue)

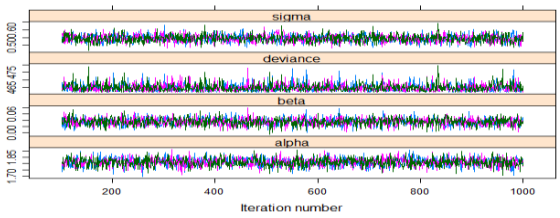

Fig. 2: Trace plots of the parameters

The above figure illustrates that on convergence to its stationary distribution, the obtained Markov chain typically looks like a random scatter about some stable mean value.

4.2.2 The following figure shows the plots of posterior densities of the parameters involved for each of the three chains. (generated using the CODA package with 1000 iterations)

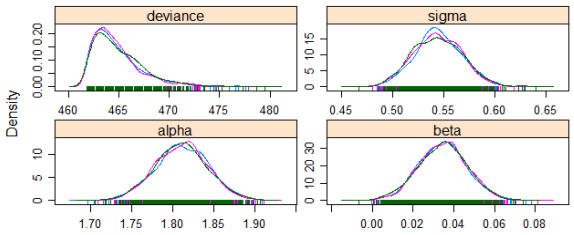

Fig. 3: Plot of the posterior densities of the parameters for each of the three chains

It is observed from the Fig. 3 that the posterior estimates of the parameters are almost normally distributed.

4.3 PREDICTION SUMMARY

The simulation-based estimate of the predictive value of the outcome variable is obtained. The resulting predicted value of the future outcome variable is computed for a particular value of BMI as the $(n+1)^{ths}$ value in the dataset. As the first case, the natural log of the Median Liver Stiffness (outcome variable) comes out to be 1.7303 for the given value of BMI as 25. The 95% credible intervals of the predicted mean value (1.65, 1.81) and predicted value (0.62, 2.79) for the given value of BMI = 25 are also obtained by simulation in R. Similarly, we can generate predictive mean values and predicted values of Median Liver Stiffness for any value of BMI.

V. SIMULATION STUDY

Computational algorithms of MCMC method are employed for obtaining numeric results. The Gibbs sampler **[6]** is one of the most widely used algorithms for simulating Markov chains. Thus, here, we perform Gibbs sampling using R2OpenBUGS package in R, a class of sampling algorithms in MCMC method and run an MCMC simulation to get estimates for the desired unknown parameters. The Deviance Information Criterion (DIC) **[7]** is the posterior mean of deviance plus the pD, the estimated effective number of parameters in the posterior distribution. Also, the potential scale reduction factor Rhat, indicates a good mixing of the three chains and thus approximate convergence gives the estimate of expected predicted error. The predicted value for the future outcome variable has been also obtained by extending the simulation of the model and accordingly, DIC for the model is also worked out. Further, Bayesian HPD credible interval, with probability 0.95 has been also computed for the estimates.

VI. INFERENCE FROM THE BUGS MODEL

6.1 THE POSTERIOR ESTIMATES AND THE CORRESPONDING HPD INTERVAL OF THE PARAMETERS

The number of chains to be run (n.chain = 3) and the number of iterations (n.iter = 10000) for each chain are specified. For each parameter, the summary statistics such as mean, standard deviation, its quantiles and convergence diagnostics are obtained.

Deviance is the general measure of model adequacy. In most of the situations, many authors have suggested using the posterior mean deviance D = E[D] as a measure of fit. In our model the posterior mean deviance comes out to be 464.8

Table 5: Posterior estimates and the Highest Posterior Density Interval of the parameters

	Mean	*sd*	*25%*	*50%*	*97.5%*	*Rhat*	***HPD Interval (0.95)***
alpha	**1.810**	0.032	1.789	1.810	1.874	1.00094	**(1.74,1.86)**
beta	**0.035**	0.011	0.027	0.035	0.058	1.00092	**(0.012,0.05)**
sigma	**0.544**	0.023	0.528	0.543	0.591	1.001	**(0.498,0.588)**
Deviance	**464.806**	2.465	463	464.2	471.102	1.001	**(461.8,469.6)**
pD=3							
DIC=467.8							

Thinning interval = 1, Sample size per chain = 9900

From this table the posterior estimates of alpha, beta and sigma are as follows: **alpha= 1.81 (0.03), beta= 0.03 (0.01) and sigma= 0.54 (0.02)**

Highest Posterior Density interval (HPD) **[8]** is also computed and comes out to be the shortest interval among all the Bayesian credible intervals. It clearly indicates that the posterior estimates of alpha, beta, sigma obtained from Table 4 are highly supported in terms of the Highest Posterior Density interval (HPD) with Bayesian Credibility.

6.2 POSTERIOR DENSITIES OF THE PARAMETERS

Posterior density of the parameter helps to draw inferences of unknown quantity of interest. The simulation is performed and the resulting posterior density plots of the parameters alpha, beta and sigma are as follows:

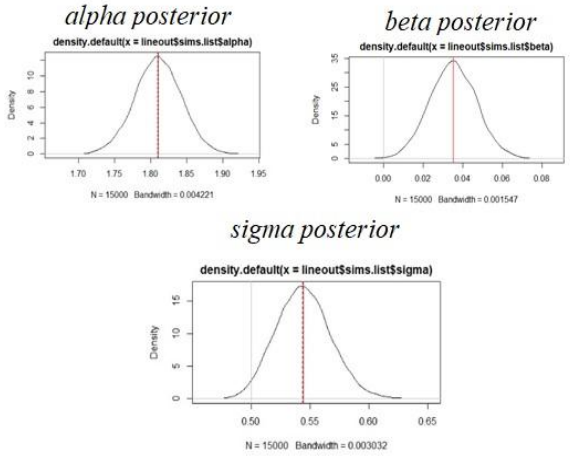

Fig 4: Plots of the posterior densities of the parameters alpha, beta and sigma.

Posterior density estimates of the parameters alpha, beta, sigma resemble that of a normal distribution with the red dotted line representing the mean of the respective distributions.

6.3 PREDICTED VALUE OF OUTCOME VARIABLE (MEDIAN LIVER STIFFNESS) FOR GIVEN VALUES OF EXPLANATORY VARIABLE (BMI)

Different values are taken for BMI =25, 26, 26.8, 27 and independent prediction of future outcome value y* are obtained for each value of BMI. After every prediction, the dataset is revised and the associated DIC's are calculated for the model.

Table 6: Predicted values of Median Stiffness for different values of BMI with their respective DIC's

PREDICTED VALUE OF MEDIAN STIFFNESS (z_i)	PREDICTED VALUE y_i* (95%Credible Interval)	BMI x_i*	EXPECTED PREDICTIVE ERROR (DIC)	MEAN (95%Credible Interval)
5.642	1.730 (0.622, 2.795)	25	**467.8**	1.735 (1.65, 1.81)
5.856	1.767 (0.669, 2.836)	26	**467.8**	1.772 (1.705, 1.84)
6.032	1.797 (0.697, 2.867)	26.8	**467.8**	1.802 (1.739, 1.866)
6.073	1.804 (0.703, 2.875)	27	**467.8**	1.809 (1.747, 1.874)

From the above table, we observe that the DIC remains constant which signifies model accuracy and the model stands out to be appropriate for further prediction. It will not be out of place to mention here that the Model is Robust in terms of DIC.

6.4 POSTERIOR PREDICTIVE PLOTS

6.4.1 The posterior predictive distribution **[9]** of the outcome variable using simulation is computed.

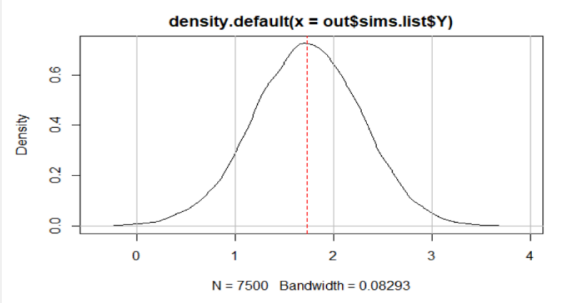

Fig. 5: Posterior predictive distribution plot of response variable for explanatory variable= 25

The above figure indicates that the posterior predictive plot of the outcome variable is approximately a normal distribution where the red dotted line indicates the mean of the distribution.

6.4.2 The underlying plot illustrates the predictive density plot of the mean of the predicted distribution taking N= 2160000

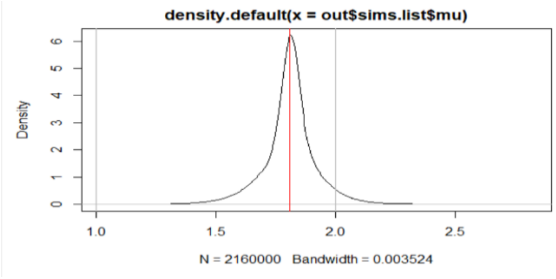

Fig. 6: Posterior predictive distribution plot of the mean of predicted distribution.

6.4.3 For the robustness of the model, posterior predictive density of the outcome variable for different values of explanatory variable (BMI) as 25, 26, 26.8, 27 is shown below.

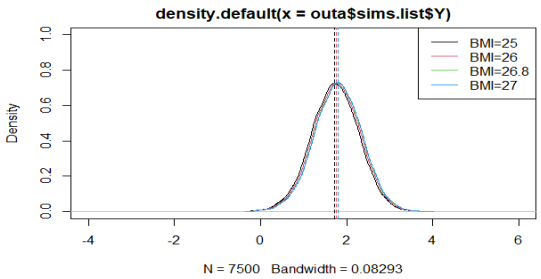

Fig. 7: Posterior predictive distribution plots of response variable for different values of explanatory variable

From the above plot, we observe that each distribution plot is almost same which shows the model is robust with respect to DIC.

VII. CONCLUSION

DIC is a hierarchical modelling generalization of the Akaike information criterion (AIC). Thus, a Bayesian Simple linear regression model is developed under the violation of normality assumptions for the error terms. The outcome variable "Median Stiffness of the liver" is predicted for different values of given explanatory variable "BMI" in this case. Since the DIC of the Bayesian model remains constant and does not increase throughout while predicting the Median Stiffness for any other value of BMI, it clearly indicates model accuracy and shows the robustness of the model in terms of DIC. Thus, the proposed Bayesian prediction model is recommended for the prediction of an outcome

variable under the violation of normality assumption.

ACKNOWLEDGEMENTS

We are grateful to Dr. A. K. Tewari, MD (Med), Dip. Diab (UK), Diabetes Care Centre, Lucknow for his invaluable assistance and for providing the dataset. His generosity and expertise improved this study in innumerable ways.

REFERENCES

[1]. Zellner Arnold, Bayesian and Non-Bayesian Analysis of the Log-Normal Distribution and Log-Normal Regression, *Journal of the American Statistical Association*, Jun., 1971, *Vol. 66, No. 334* (Jun., 1971), PP. 327-330

[2]. David Lunn, Christopher Jackson, Nicky Best, Andrew Thomas, David Spiegelhalter, *A Practical Introduction to Bayesian Analysis* (CR Press, U.S. ,2013) PP 84-87

[3]. Edward Greenberg. *Introduction to Bayesian Econometrics* (Cambridge University Press, New York, 2008) Appendix B PP 192-193

[4]. Sturtz, S., Ligges, U., Gelman, A. (2005): *R2WinBUGS: A Package for Running WinBUGS from R*. Journal of Statistical Software *12(3)*, 1-16.

[5]. Best, N. G., Cowles, M. K., and Vines, S. K. *CODA: Convergence Diagnosis and Output Analysis software for Gibbs Sampler output*: Version 0.3(1995). Medical Research Council Biostatistics Unit, Cambridge, UK

[6]. Casella, G. and George, E. I, Explaining the Gibbs sampler *The American Statistician* (1992) *46*, 167–74.

[7]. David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, Angelika van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*. B (2002) 64, Part 4, pp. 583–63

[8]. Chen, M. H. & Shao, Q. M. Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics* (1999) *8(1)*:69–92.

[9]. Gelman A, Carlin J, Stern H, Rubin D *Bayesian Data Analysis*. CRC Press, Boca Raton,2 edition,2003)