Bayesian Unit-Lindley Model: Applications to Gasoline Yield and Risk Assessment Data

Weerinrada Wongrin[1], Sakuna Srianomai[2], and Yuwadee Klomwises[2]*

[1]Department of Statistics, Faculty of Science, Chiang Mai University, Mueang Chiang Mai, Chiang Mai, 50200

[2]Department of Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Ladkrabang, Bangkok, 10520

* Corresponding author. E-mail address: yuwadee.kl@kmitl.ac.th

Received: 27 June 2019; Revised: 20 November 2019; Accepted: 9 December 2019

Abstract

The regression model for the response variable with bounded domain is discussed. The baseline distribution called the unit Lindley distribution is considered. In the context of regression structure, the logit function is utilized with the unit Lindley model. Then, we have developed the Bayesian unit Lindley regression based on a frequently used prior. Additionally, we also investigate the specific prior for all standardized exploratory variables. The syntax of JAGS for the proposed model is included. In application study, the Bayesian unit Lindley regression is applied to two different datasets where response variables are associated with gasoline yield and risk assessment respectively. Based on the result of estimates and log-likelihood values, it is important to point out that the Bayesian unit-Lindley regression can improve the performance of the classical one.

Keywords: rate data, Bayesian, Unit distribution, gasoline yield, risk assessment

Introduction

The Lindley distribution have been studied and applied to wildly kind of lifetime data. Some applications include waiting times before service of bank customers (Ghitany, Atieh, & Nadarajah, 2008), lifetime of patients with squamous cell carcinoma (Mazucheli & Achcar, 2011), survival times of guinea pigs infected with virulent tubercle bacilli (Shanker, Sharma, & Shanker, 2013), times to breakdown of an insulating fluid (Raqab, Al-Jarallah, & Al-Mutairi, 2017). Recently, Mazucheli, Menezes, and Chakraborty (2019) have extended the Lindley distribution to model data on unit interval, in particular rate and proportion. Specifically, they have developed the one parameter unit-Lindley distribution with bounded domain as an alternative to beta distribution. By applied suitable parameterization and logit link function, they also proposed regression model for continuous bounded data and the interpretation of regression parameters is straightforwardly related to the mean of a dependent variable. In application study, they created regression model by performing maximum likelihood estimation. Moreover, they concluded that the unit-Lindley regression model fits better than the Beta regression for proportion of households with inadequate water supply.

As mentioned above, the unit Lindley regression model can be considered as an alternative model to analyze rate and proportion data. In this work, we have been motivated to propose regression model based on the unit Lindley distribution (Mazucheli et al., 2019). By considering regression coefficients to be random variables rather than constant, the Bayesian unit-Lindley model with informative prior was obtained. Besides, the use of informative prior can be perceived as adding a number of observations to a given sample size (Ali, Aslam, & Kazmi, 2013). The rest of this work are organized as follows. The unit Lindley distribution and its properties are discussed. The generalized linear model based on unit Lindley response variable is described together with its maximum likelihood estimation. Next, the Bayesian unit Lindley model is created. Finally, application study is performed to compare performances between traditional unit Lindley regression model and Bayesian unit Lindley regression model.

Unit-Lindley Linear Model

Mazucheli et al. (2019) proposed the unit Lindley distribution with the cumulative distribution function (cdf) and the probability density function (pdf) are respectively

$$F(y \mid \theta) = 1 - \left(1 + \frac{\theta y}{(1+\theta)(x-1)}\right)\exp\left(-\frac{\theta x}{1-x}\right)$$

and

$$f(y \mid \theta) = \frac{\theta^2}{1+\theta}\left(1+x\right)^{-3}\exp\left(-\frac{\theta x}{1-x}\right)$$

where $0 < y < 1$ and $\theta > 0$.

The unit Lindley distribution is a strongly unimodal distribution with exponential tail. It has the exist moments, and some properties of the unit Lindley distribution such as the hazard rate function, mean residual life function, and mean deviation were provided in Mazucheli et al. (2019).

They also developed linear model based on the unit-Lindley distribution. First, they reparametrized the unit-Lindley probability density function in terms of μ_i as

$$\mathrm{f}\left(y_i | \mu_i\right) = \frac{\left(1-\mu_i\right)^2}{\mu_i\left(1-y_i\right)^3}\exp\left(-\frac{y_i\left(1-\mu_i\right)}{\mu_i\left(1-y_i\right)}\right)$$

where $i = 1, \ldots, n$. Let $\mathbf{x}_i^T = (1, x_{i1}, \ldots, x_{ip})^T$ is a vector of covariates and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is a vector of regression coefficient. Another component of the unit-Lindley linear model is a link function:

$$g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$$

It maps the unbounded space of the linear predictor into bounded sample space (Smithson & Verkuilen, 2006). Specifically, bounded sample space of unit-Lindley dependent random variable ~~of~~ is on interval (0,1). In addition, there are many forms of link function can be used, such as the probit or complementary log-log link, as well as any cumulative distribution function corresponding to a continuous distribution. Referring to the work of Mazucheli et al. (2019), we also use the logit link function written as

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \mathbf{x}_i^T\boldsymbol{\beta}$$

By inverting the logit link function, ones can get predicted values by following equation

$$\mu_i = \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}.$$

Maximum Likelihood Estimation

For the sake of parameter estimation, the estimates of $\boldsymbol{\beta}$ can be obtained by maximizing the likelihood or log-likelihood function. Let response variable $Y_1,\ldots,Y_n$ have the unit Lindley distribution denoted by $Y_i \sim \mathrm{UL}(\mu_i)$, $i=1,\ldots,n$ and $\mathbf{x}_i^T=(1,x_{i1},\ldots,x_{ip})^T$ be a vector of covariate based on for the i-th observation. Then, the likelihood and the log likelihood functions with covariates can be represented respectively as

$$L(\boldsymbol{\beta})=\prod_{i=1}^{n}\frac{\left(\dfrac{1}{1+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}\right)^2}{\left(\dfrac{\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}\right)(1-y_i)^3}\exp\left(-\frac{y_i}{(1-y_i)\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}\right)$$

and

$$\ell(\boldsymbol{\beta})=2\sum_{i=1}^{n}\log\left(\frac{1}{1+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}\right)-\sum_{i=1}^{n}\log\left(\frac{\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}\right)$$
$$-3\sum_{i=1}^{n}\log(1-y_i)-\sum_{i=1}^{n}\frac{y_i}{1-y_i}\left(\frac{1}{\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}\right)$$

Moreover, numerically maximized of likelihood and log-likelihood function can be accomplished by optim function in R (R Core Team, 2018).

Bayesian Unit-Lindley Linear Model

In previous section, the traditional approach for parameter estimation is discussed. On the other hand, in this section, we introduce Bayesian framework for unit-Lindley regression model. The Bayesian unit-Lindley linear model mainly created based on the likelihood function, prior distribution, and posterior distribution, denoted respectively by $L(\boldsymbol{\beta})$, $p(\boldsymbol{\beta})$, and $p(\boldsymbol{\beta}\mid\boldsymbol{y})$.

If a response variable $\mathbf{Y}$ is distributed as unit-Lindley and $(\mathbf{x}_i,\mathbf{y}_i)$ be regression data of size n, the Bayesian unit-Lindley regression model is

$$y_i\mid\mu_i \sim \mathrm{UL}(\mu_i),$$
$$\mu_i=g^{-1}\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)$$

and

$$\boldsymbol{\beta}\sim p(\boldsymbol{\beta})$$

Based on likelihood function and the prior distribution, the posterior distribution is written by

$$p(\boldsymbol{\beta}\mid\boldsymbol{y})=\frac{L(\boldsymbol{\beta})p(\boldsymbol{\beta})}{\int L(\boldsymbol{\beta})p(\boldsymbol{\beta})d\boldsymbol{\beta}}$$
$$\propto\prod_{\mathrm{i}=1}^{\mathrm{n}}\frac{\left(\dfrac{1}{1+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}\right)^2}{\left(\dfrac{\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}\right)(1-y_i)^3}\exp\left(-\frac{y_i}{(1-y_i)\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}\right)\times p(\boldsymbol{\beta})$$

Accordingly, under squared error loss function, Bayesian estimator of the regression coefficient will be $\mathrm{E}\left(\boldsymbol{\beta} \mid \boldsymbol{y}\right)$.

Normal prior

The prior distribution can be considered the most important component of the Bayesian inference as it represents the information about an uncertain parameter. We refer the reader to Smithson and Verkuilen (2006), Ali et al. (2013), and Ali (2015) for more details about the impact of prior distribution toward Bayesian estimator. In the context of generalized linear model, the most frequency used informative prior distributions is normal distribution (Dey, Ghosh, & Mallick, 2000). Let prior distribution for $\boldsymbol{\beta}$ be the normal distribution, denoted by $\boldsymbol{\beta} \sim \mathrm{N}(\mu_{\beta}, \sigma_{\beta}^{2})$, then

$$p\left(\boldsymbol{\beta}\right) = \frac{1}{\sqrt{2\pi\sigma_{\beta}}} \exp\left(-\frac{\left(\beta_j - \mu_{\beta}\right)}{2\sigma_{\beta}^{2}} \right)$$

There are many researchers devoted to investigate on the prior distribution and relative hyperparameters. Recently, Gelman, Jakulin, Pittau, and Su (2008) come up with an idea of defining weakly informative prior for logistic and other regression models. As independence with difference scale should not have the same prior distribution. Gelman et al. (2008) suggested that all independence variables should be standardized to have mean 0 and standard deviation 0.5 except for qualitative variable.

Posterior distribution and Gibbs Sampler

The posterior distribution will integrate the sample information from the likelihood function with accessible parameters information from the prior distribution. Subsequently, the posterior distribution with respect to the normal becomes

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) \propto \prod_{\mathrm{i}=1}^{\mathrm{n}} \frac{\left(\dfrac{1}{1+\exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)} \right)^{2}}{\left(\dfrac{\exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)} \right) \left(1-y_i\right)^{3}} \exp\left(-\frac{y_i}{\left(1-y_i\right)\exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)} \right) \times \frac{1}{\sqrt{2\pi\sigma_{\beta}^{2}}} \exp\left(-\frac{\left(\boldsymbol{\beta} - \mu_{\beta}\right)}{2\sigma_{\beta}^{2}} \right)$$

As the posterior distribution does not have an explicit form, the computational methods called Gibbs sampler, the best known MCMC sampling algorithms, was applied in order to find $\mathrm{E}\left(\boldsymbol{\beta} \mid \boldsymbol{y}\right)$. By setting some initial points, the Gibbs sampler algorithm random will walk through parameter space. The basic scheme Gibbs sampler is given as follows (Joseph et al, 2001)

Step 0. Choose as arbitrary starting point $\boldsymbol{\beta}^{(0)}$

Step 1. Generate $\boldsymbol{\beta}^{(\mathrm{i}+1)}$ as follows:

$$\text{Generate } \beta_0^{(\mathrm{i}+1)} \sim p(\beta_0 \mid \beta_1^{(\mathrm{i})}, \beta_2^{(\mathrm{i})}, \ldots, \beta_p^{(\mathrm{i})}, \boldsymbol{y});$$

$$\text{Generate } \beta_1^{(\mathrm{i}+1)} \sim p(\beta_1 \mid \beta_0^{(\mathrm{i})}, \beta_2^{(\mathrm{i})}, \ldots, \beta_p^{(\mathrm{i})}, \boldsymbol{y});$$

$$\vdots$$

$$\text{Generate } \beta_p^{(\mathrm{i}+1)} \sim p(\beta_p \mid \beta_0^{(\mathrm{i})}, \beta_2^{(\mathrm{i})}, \ldots, \beta_{p-1}^{(\mathrm{i})}, \boldsymbol{y});$$

Step 2. Set $i = i+1$ and go to step 1.

Besided, trace plot and density plot of MCMC chains will be applied to access convergence to stationarity. Then, Gelman plot from R2jags package (Su & Yajima, 2015) in R is also presented in order to examine convergence of the average (Robert, Casella, & Casella, 2010). Moreover, the JAGS syntax for Bayesian unit-Lindley regression model with logit link is presented in **Appendix.**

Application to Real Data

In this section, two real data are analyzed based on the unit-Lindley regression model and the Bayesian unit-Lindley regression model. For the Bayesian one, we use two chains based on difference initial values. In addition, each chain generates 400,000 interactions and discard the first 100,000 as burn-in. We also consider Gelman plot in Rjags package (Su and Yajima, 2015) for all parameters. Furthermore, the Gelman plots show the scale-reduction over time step. The factor of 1 means that there is no difference between chain. The density plot in coda package (Plummer et al., 2006) is also included to determine symmetric of the MCMC density. Finally, the performances of candidate models are discussed.

Gasoline data

The first data is Gasoline data consisting 32 observations. For this data, the aim of the research was to evaluate how distillation properties of crude have an impact on percentage yield of gasoline (Hand, Daly, McConway, Lunn, & Ostrowski, 1993). In addition, yield is response variable with other independent variables described as follow (Cribari-Neto & Zeileis, 2009)

yield	proportion of crude oil converted to gasoline after distillation and fractionation.
gravity	crude oil gravity (degrees API).
pressure	vapor pressure of crude oil (lbf/in2).
temp10	temperature (degrees F) at which 10 percent of crude oil has vaporized.
temp	temperature (degrees F) at which all gasoline has vaporized.

The results of estimates, standard error, and log likelihood from MLE and Bayesian are shown in Table 1.

Table 1 The results of regression coefficients estimation together with log likelihood regarding to MLE and Bayesian methods for gasoline data (n =32)

	MLE		Bayesian	
	Estimates	SE	Estimates	SE
Intercept	-1.7113	0.157	-1.6657	0.0003
β_1	4.8239	4.251	0.3660	0.0079
β_2	-3.0801	7.913	2.2829	0.0131
β_3	-7.9884	8.889	-7.1447	0.0147
β_4	13.2996	3.665	13.5582	0.0066
log likelihood	32.1818		33.0260	

Based on greater log likelihood values, the Bayesian unit-Lindley regression model is more appropriate to this data than the unit-Lindley regression model. Therefore, the regression structure of this data should be

$$\text{logit}(\mu_i) = -1.6657 - 0.366\text{gravity}_i + 2.2829\ \text{pressure}_i - 7.1447\ \text{temp10}_i + 13.5582\text{temp}_i$$

For Bayesian unit-Lindley regression model, the Gelman plots with respect to β_0, β_1, β_2, β_3, and β_4 are demonstrated respectively. As shown in Figure 1, all regression coefficients of the unit-Lindley regression model, the factors close to 1, therefore, the MCMC chains are similar.

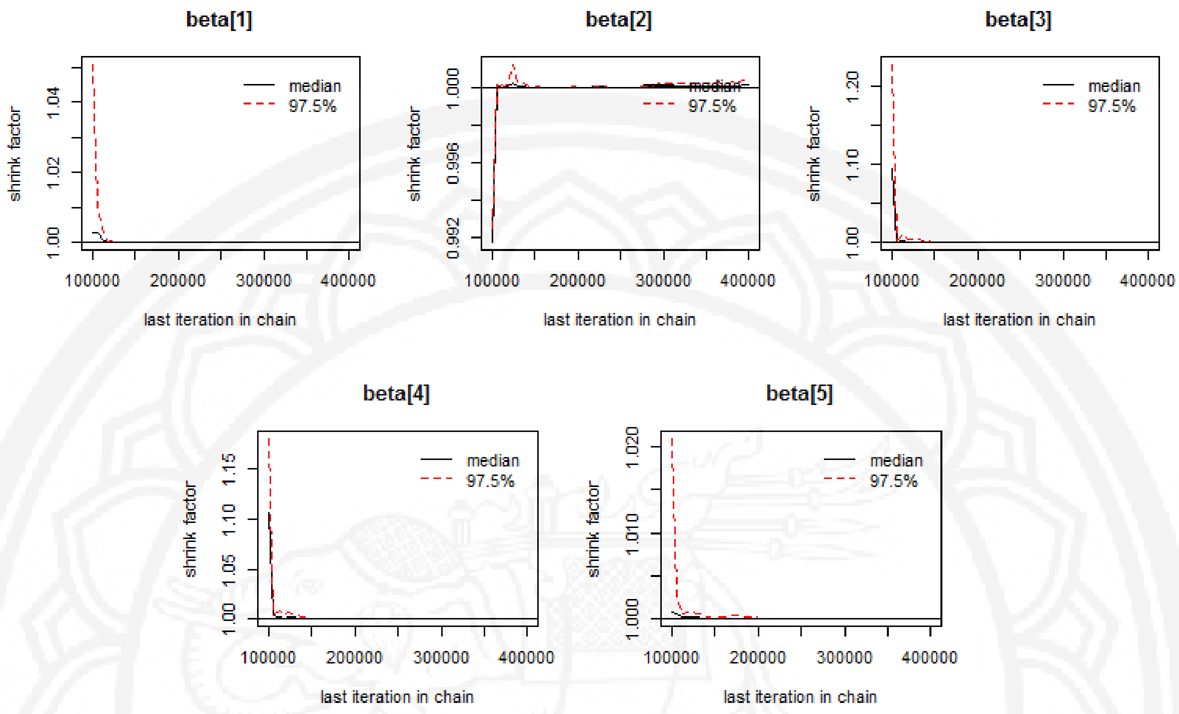

Figure 1 Gelman plots of MCMC chains related to β_0, β_1, β_2, β_3, and β_4 respectively for gasoline data

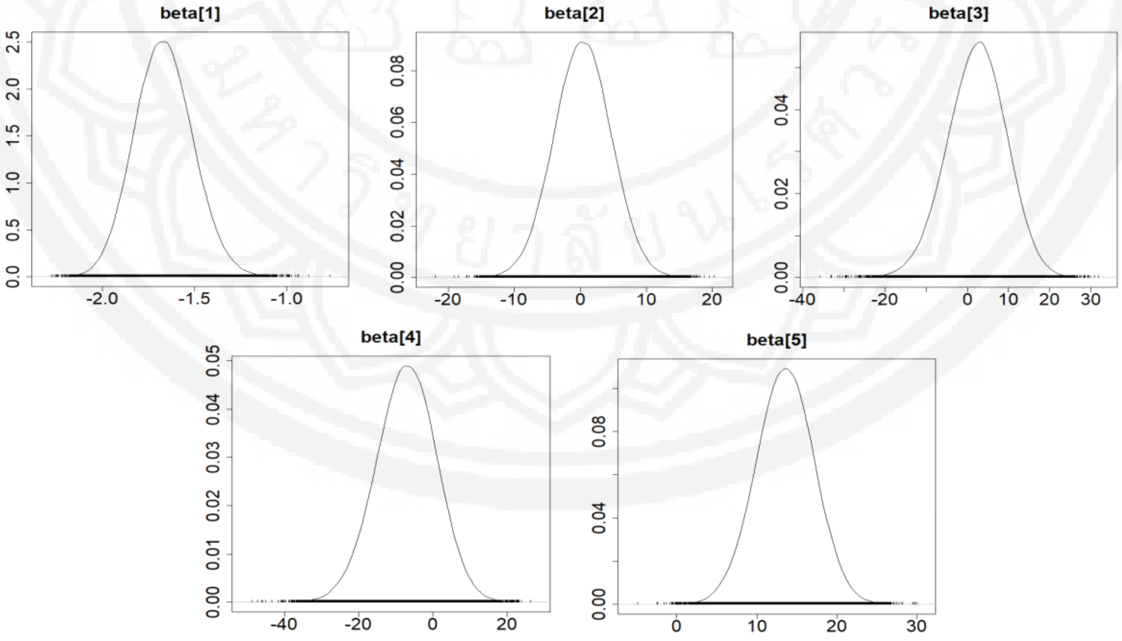

Figure 2 Density plots of MCMC chains related to β_0, β_1, β_2, β_3, and β_4 respectively for gasoline data

Risk assessment data

The second data is related to risk managers cost effectiveness (Schmit & Roth, 1990). It can be obtained from the personal web page of Professor E. Frees (https:// sites.google.com/a/wisc.edu/jed-frees/). The objective of the study was to investigate theeffctiveness of risk management while controlling for organizational risk characteristics. This data consists of 7 variables where FIRMCOST can be regarded as response variable and others are exploratory variables as follows

FIRMCOST	total property and casualty premiums and uninsured losses as a percentage of total assets (divided by 100)
ASSUME	Per occurrence retention amount as a percentage of total assets
CAP	Indicates that the firm owns a captive insurance company
SIZELOG	Logarithm of total assets
INDCOST	A measure of the firm's industry risk
CENTRAL	A measure of the importance of the local managers in choosing the amount of risk to be retained
SOPH	A measure of the degree of importance in using analytical tools.

By applying two regression model to risk assessment, the estimates, standard error, and log likelihood are obtained and presented in Table 2.

Table 2 The results of regression coefficients estimation together with log likelihood regarding to MLE and Bayesian methods for risk managers cost effectiveness data (n =73)

	MLE		Bayesian	
	Estimates	SE	Estimates	SE
Intercept	-2.122	0.1673	-2.281	0.00028
β_1	1.015	3.9709	-9.528	0.00391
β_2	2.155	0.3076	1.891	0.00059
β_3	-3.835	2.0022	-12.68	0.00409
β_4	2.458	2.6014	19.348	0.00656
β_5	6.008	2.7650	2.278	0.00473
β_6	-8.813	3.0619	-8.906	0.00556
log likelihood	21.6301		45.6233	

Referring to log likelihood values in Table 2, we can conclude that the Bayesian unit-Lindley regression model is more suitable to this data that the unit-Lindley regression model. In conclusion, the regression model for risk managers cost effectiveness is

$$\begin{aligned}\operatorname{logit}(\mu_i) &= -2.281 - 9.528\text{ASSUME}_i + 1.891\text{ CAP}_i - 12.68\text{ SIZELOG}_i + 19.348\text{ INDCOST}_i \\ &+2.278\text{ CENTRAL}_i - 8.906\text{ SOPH}_i\end{aligned}$$

In addition, the convergence diagnostics of Gelman and Rubin with respect to β_0, β_1, β_2, β_3, β_4, β_5, and β_6 are illustrated in Figure 3. The factors of all parameters are near 1, which indicates that each MCMC chains is convergent.

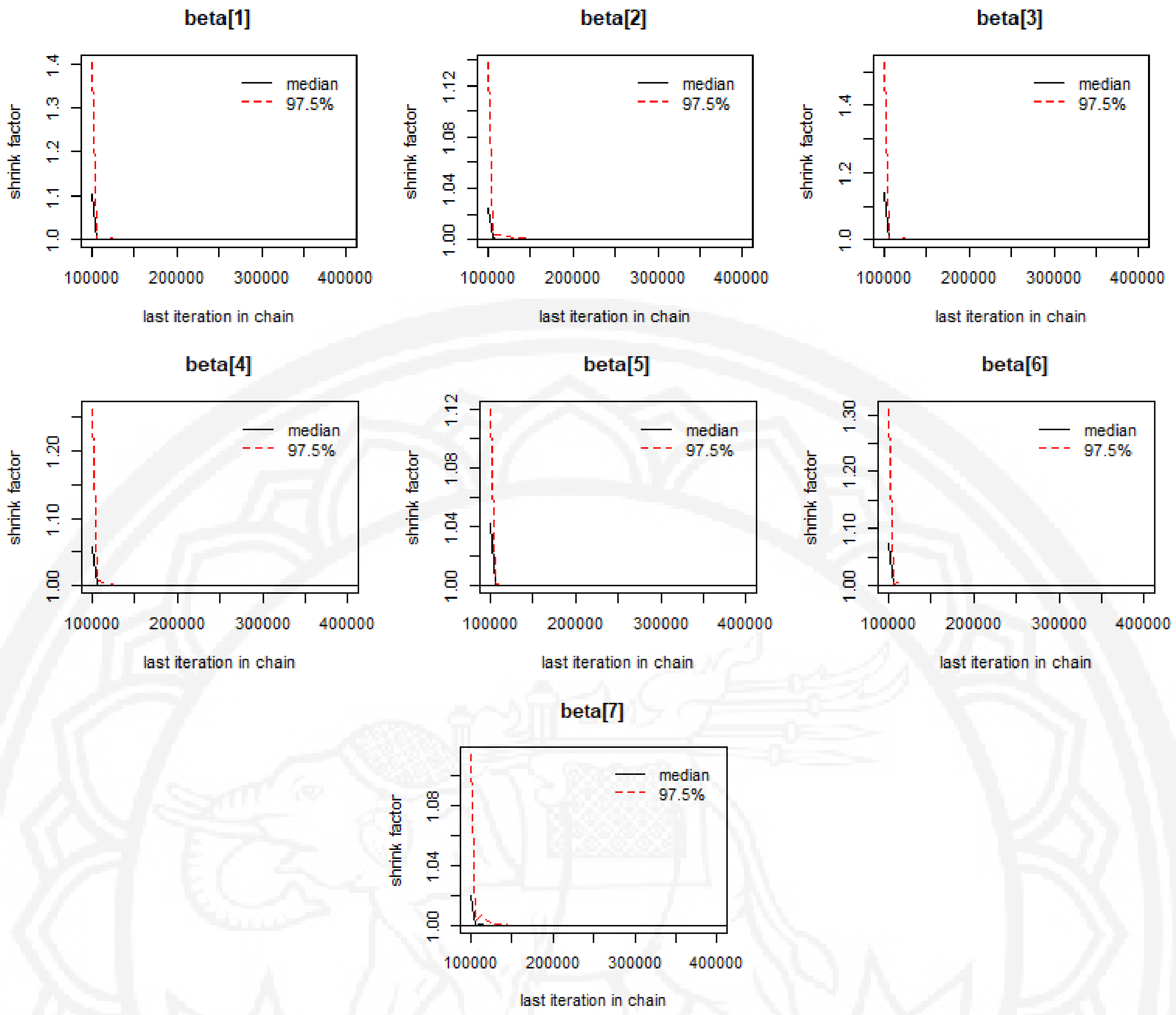

Figure 3 Gelman plots of MCMC chains related to β_0, β_1, β_2, β_3, β_4, β_5, and β_6 respectively for risk assessment data

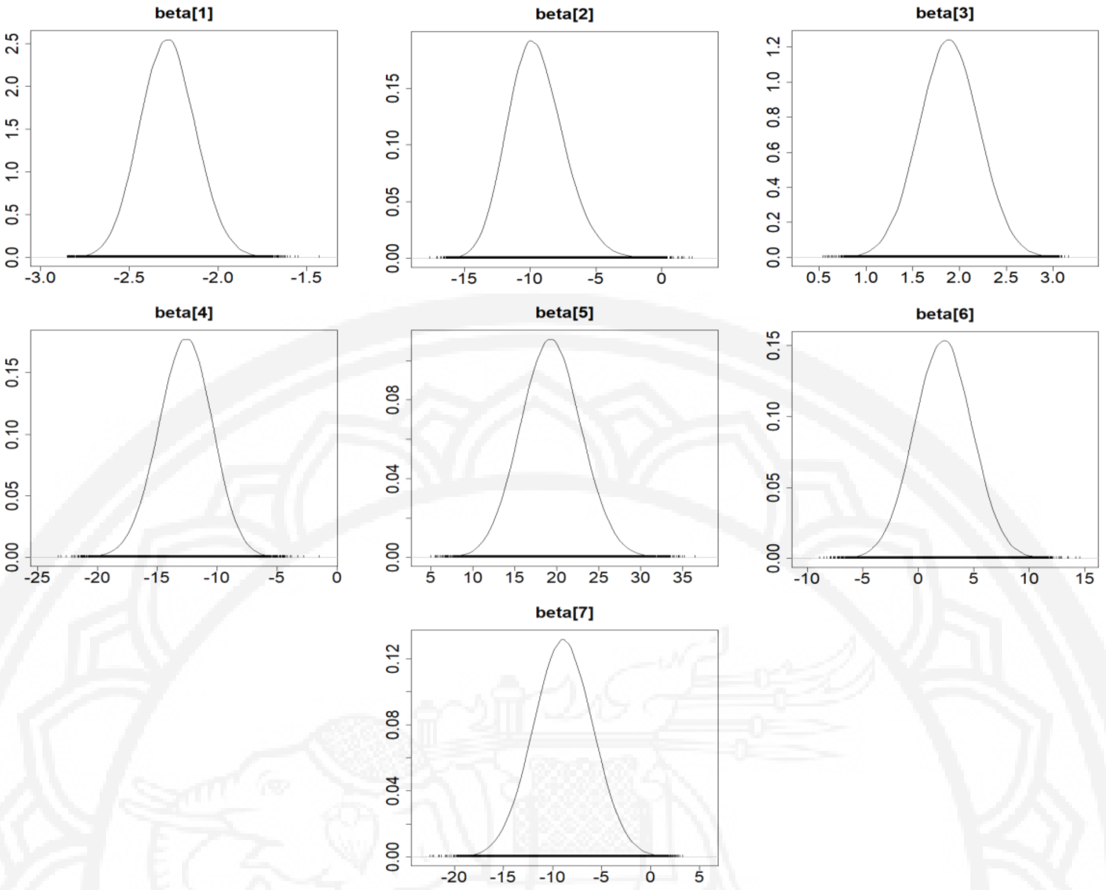

Figure 4 Density plots of MCMC chains related to β_0, β_1, β_2, β_3, β_4, β_5, and β_6 respectively for risk assessment data

Conclusions

The use of Bayesian framework could be recommended to the regression model with bounded response variable on (0,1). As there are many authors have suggested the use of Bayesian estimator to linear model such as Branscum, Johnson, and Thurmond (2007), Gelman et al. (2008). In this work, the Bayesian unit-Lindley regression is created. The Normal prior and standardization of exploratory variable is discussed. For comparison purposes, the traditional and Bayesian unit-Lindley regression models are applied to two real data. There are Gasoline data and Risk assessment data where response variable are gasoline yield and percentage of total asset respectively. Remarkably, the Bayesian unit Lindley regression show better results than the unit Lindley model in term of log-likelihood values.

Appendix

The JAGS syntax for the Bayesian unit-Lindley regression model is

```
BLindley_reg<-function(){
   for (i in 1:n){
   zeros[i] ~ dpois(phi[i])
   phi[i] <- - l[i] + C
   l[i] <- 2*log(1-mu[i])-log(mu[i])-3*log(1-y[i])-(y[i]*(1-
   mu[i]))/(mu[i]*(1-y[i]))
   logit(mu[i]) <- inprod(X[i,], beta[])
   }
   #prior
   for (j in 1:J){
   beta[j] ~ dnorm(0,0.001)
   }
   }
```

References

Ali, S. (2015). On the Bayesian estimation of the weighted Lindley distribution. *Journal of Statistical Computation and Simulation*, *85*, 855-880. https://doi.org/10.1080/00949655.2013 .847442

Ali, S., Aslam, M., & Kazmi, S. M. A. (2013). A study of the effect of the loss function on Bayes Estimate, posterior risk and hazard function for Lindley distribution. *Applied Mathematical Modelling*, *37*, 6068-6078. https://doi.org/10.1016/j.apm.2012.12.008

Branscum, A. J., Johnson, W. O., & Thurmond, M. C. (2007). Bayesian beta regression: applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian & New Zealand Journal of Statistics*, *49*(3), 287-301. https://doi.org/10.1111/j.1467-842X.2007.00481.x

Cribari-Neto, F., & Zeileis, A. (2009). Beta regression in R. *Journal of Statistical Software*, *34*, 1-24. http://dx.doi.org/10.18637/jss.v034.i02

Dey, D. K., Ghosh, S. K., & Mallick, B. K. (2000). *Generalized linear models:A Bayesian perspective*. New York, USA: CRC Press.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*, 1360-1383. https://doi.org/10.1214/08-AOAS191

Ghitany, M. E., Atieh, B., & Nadarajah, S. (2008). Lindley distribution and its application. *Mathematics and computers in simulation*, *78*, 493-506. https://doi.org/10.1016/j.matcom.2007.06.007

Hand, D. J., Daly, F., McConway, K., Lunn, D., & Ostrowski, E. (1993). *A handbook of small data sets*. New York, USA: CRC Press.

Mazucheli, J., & Achcar, J. A. (2011). The Lindley distribution applied to competing risks lifetime data. *Computer methods and programs in biomedicine*, *104*, 188-192.

Mazucheli, J., Menezes, A. F. B., & Chakraborty, S. (2019). On the one parameter unit-Lindley distribution and its associated regression model for proportion data. *Journal of Applied Statistics*, *46*, 700-714. https://doi.org/10.1080/02664763.2018.1511774

Plummer, M., Best, N., Cowles, Kate., & Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, *6*, 7-11.

R Core Team. (2018). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* Retrieved from https://www.R-project.org/

Raqab, M. Z., Al-Jarallah, R. A., & Al-Mutairi, D. K. (2017). Closeness of Lindley distribution to Weibull and gamma distributions. *Communications for Statistical Applications and Methods*, *24*, 129-142. https://doi.org/10.5351/CSAM.2017.24.2.129

Robert, C. P., Casella, G., & Casella, G. (2010). *Introducing Monte Carlo methods with R*. New York, USA: Springer.

Schmit, J. T., & Roth, K. (1990). Cost effectiveness of risk management practices. *Journal of Risk and Insurance*, *57*, 455-470. http://doi.org/10.2307/252842

Shanker, R., Sharma, S., & Shanker, R. (2013). A two-parameter Lindley distribution for modeling waiting and survival times data. *Applied Mathematics*, *4*, 363-368. http://dx.doi.org/10.4236/am.2013.42056

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, *11*, 54-71. https://doi.org/10.1037/1082-989X.11.1.54

Su, Y. S., & Yajima, M. (2015). *R2jags: Using R to Run 'JAGS'. R package version 0.5-7.* Retrieved from https://CRAN.R-project.org/package=R2jags