World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

A new Bayesian ridge estimator for logistic regression in the presence of multicollinearity

Folashade Adeola Bolarinwa [1,*], Olusola Samuel Makinde [2] and Olusoga Akin Fasoranbaku [2]

[1] *Department of Statistics, Federal Polytechnic, Ado-Ekiti, Nigeria.*
[2] *Department of Statistics, Federal University of Technology, Akure, Nigeria.*

World Journal of Advanced Research and Reviews, 2023, 20(03), 458–465

Publication history: Received on 19 October 2023; revised on 01 December 2023; accepted on 04 December 2023

Article DOI: https://doi.org/10.30574/wjarr.2023.20.3.2415

Abstract

This research introduces the Bayesian schemes for estimating logistic regression parameters in the presence of multicollinearity. The Bayesian schemes involve the introduction of a prior together with the likelihood which resulted in the posterior distribution that is not tractable, hence the use of a numerical method i.e Gibbs sampler. Different levels of multicollinearity were chosen to be $\rho = 0.80,0.85,0.90,0.95,0.99\ and 0.999$ to accommodate severe, very severe and nearly perfect state of multicollinearity with sample sizes taken as 10,20,30,50,100,200,300 and 500.Different ridge parameters k were introduced to remedy the effect of multicollinearity .The explanatory variables used were 3 and 7. Model estimation was carried out using Bayesian approach via the Gibbs sampler of Markov Chain Monte Carlo Simulation. The means square error MSE of Bayesian logistic regression estimation was compared with the frequentist methods of the estimation. The result shows a minimum mean square error with the Bayesian scheme compared to the frequentist method.

Keywords: Bayesian; Logistic Regression; Multicollinearity; Mean Square Error

1. Introduction

Generalized Linear Models (GLMs) to which logistic regression belongs are a class of statistical models used for modeling the relationship between a dependent variable (response variable) and one or more independent variables (predictor variables or features). They are an extension of the traditional linear regression models and are particularly useful when dealing with the non-normal distribution of data or when the relationship between variables is not strictly linear. Like in linear regression, GLMs start with a linear predictor, which is a linear combination of the predictor variables. However, unlike linear regression, GLMs don't assume a linear relationship between the predictors and the response. The linear predictor is often denoted as η (eta).

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k. \quad (1)$$

Here, β_0, β_1, β_2, etc., are the coefficients to be estimated, and x_1, x_2, etc., are the predictor variables. GLMs introduce a link function (g) that relates the expected value of the response variable to the linear predictor η.

Logistic Regression is a statistical model used for binary classification problems, where the outcome variable is categorical and has only two possible values, often labeled as 0 and 1 (or "negative" and "positive"). It's called "logistic" because it uses the logistic function (also known as the sigmoid function) to model the probability of the binary outcome. Logistic Regression is a generalized linear model (GLM) and is widely used in various fields, including machine learning, statistics, and epidemiology.

* Corresponding author: Folashade Adeola Bolarinwa

Logistic Function: The logistic function, denoted as σ(η), is used to model the probability that the dependent variable takes the value 1. The logistic function has an S-shaped curve and is defined as:

$$\sigma(\eta) = \frac{1}{(1+e^{(-\eta)})} \quad (2)$$

Here, σ(η) represents the probability of the positive class (1), η is the linear combination of predictor variables (similar to the linear predictor in GLMs), and e is the base of the natural logarithm

Multicollinearity is defined as a statistical phenomenon that occurs when two or more predictor variables in a regression model are highly correlated with each other. In other words, it is a condition in which there is a strong linear relationship between two or more independent variables in a regression analysis.

This can cause issues in regression analysis because it violates the assumptions of the ordinary least squares (OLS) method, which is commonly used to estimate the parameters of a regression model. When multicollinearity is present, it becomes difficult to determine the individual effects of the correlated variables on the dependent variable. Several authors have develop different estimators to solve the problem of multicollinearity.(Dawodu,2020)

Adepoju and Ojo (2018) provided another estimator which is alternative to ordinary least square when multicollinearity is almost perfect. If multicollinearity is found to be present, there are several strategies to address it, such as removing one of the correlated variables, transforming the variables, or using dimensionality reduction techniques like principal component analysis (PCA).), or incorporating regularization methods such as ridge regression or LASSO (Least Absolute Shrinkage and Selection Operator) to mitigate the effects of multicollinearity as said by.(Park,T and Casella,G 2008). Hans, C. et al.(2010). explores the Bayesian Lasso in the context of survival analysis.To address multicollinearity in GLMs, similar strategies can be applied as in linear regression..

1.1. Bayesian methods of solving multicollinearity

Bayesian methods provide an alternative approach to addressing multicollinearity in regression models. They offer a framework that incorporates prior knowledge and uncertainty about the parameters and allows for more flexible modeling. Emenyonu and Mohd (2019) introduced Bayesian approach to logistic regression via markov chain monte carlo algorithm for posterior distribution to be obtained with the discovery that non-flat prior yielded a better model than the maximum likelihood estimate and the Bayesian with the non-informative flat prior.

1.2. Bayesian Logistic Regression

Is a variation of logistic regression that incorporates Bayesian principles for estimating the model parameters and making probabilistic inferences about them. Unlike traditional (frequentist) logistic regression, which uses maximum likelihood estimation (MLE) to estimate the parameters, Bayesian Logistic Regression provides a probability distribution over the parameters themselves. This makes it possible to express uncertainty about the parameter estimates and to perform Bayesian model selection and hypothesis testing.

Gelman, A et al(2013) in his book discusses regularization techniques and hierarchical modeling in Bayesian statistical methods, which can be relevant for addressing multicollinearity.

Harrison, X. A.et al. (2018). focuses on hierarchical modeling, which can be particularly useful for handling multicollinearity in complex models. This paper seeks to use the Bayesian approach to solve the problem of multicollinearity in logistic regression bringing in some existing ridge parameter of solving multicollinearity ,

2. Materials and Methods

2.1. Prior Distribution

In Bayesian Logistic Regression, a prior distribution for the model parameters is obtained. This prior reflects the beliefs about the parameters before observing any data. It encapsulates any prior knowledge or assumptions about the values. We assume a normal prior on β.

$$\beta_{j \sim N(\mu_j, \sigma^2{}_j)} \quad (3)$$

2.2. Likelihood

Similar to traditional logistic regression, Bayesian Logistic Regression uses a likelihood function that models the probability of observing the data given the model parameters. For binary classification, the likelihood is typically the binomial likelihood.

$$\text{Likelihood} = \pi(x_i)^{y_i}\big(1-\pi(x_i)\big)^{1-y_i} \tag{4}$$

where $\pi(x_i)$represents the probability of the event for i with covariate vector x_i $and y_i$ indicates the presence of $y_i = 1$, or absence y=0 of the event i.From the classical logistic regression $\pi(x_i)$ is given by:

$$\pi(x_i) = \frac{e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}}{1+e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}} \tag{5}$$

In effect the likelihood contribution from ith subject is

$$\text{Likelihood} = \left[\frac{e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}}{1+e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}}\right]^{y_i}\left[1-\frac{e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}}{1+e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}}\right]^{(1-y_i)} \tag{6}$$

Given that individual subjects are assumed independent from each other, the likelihood function over a data set of n subject is then

$$\text{Likelihood} = \prod_{i=1}^{n}\left[\left[\frac{e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}}{1+e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}}\right]^{y_i}\left[1-\frac{e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}}{1+e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}}\right]^{(1-y_i)}\right] \tag{7}$$

2.3. Posterior Distribution

The goal of Bayesian Logistic Regression is to compute the posterior distribution over the model parameters. This posterior distribution represents the updated beliefs about the parameters after observing the data. It is proportional to the product of the prior distribution and the likelihood function.

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

$$\text{Posterior} = \prod_{i=1}^{n}\left[\left[\frac{e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}}{1+e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}}\right]^{y_i}\left[1-\frac{e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}}{1+e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}}\right]^{(1-y_i)}\right]$$

$$\text{x}\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma_j}}exp\left\{-\frac{1}{2}\left(\frac{\beta_j-\mu_j}{\sigma_j}\right)^2\right\} \tag{8}$$

Bayesian Logistic Regression offers a more principled and flexible approach to logistic regression modeling, especially when dealing with small sample sizes or when prior information is available. It provides a richer understanding of parameter uncertainty and allows for more comprehensive probabilistic inference. However, it typically requires more computational resources and expertise in Bayesian methods compared to traditional logistic regression.

Bayesian methods allow for the estimation of these hyperparameters as well, which can lead to more robust model tuning.

Logistic Regression Using Pólya-Gamma Latent Variables which this paper seeks to apply in the presence of multicollinearity. Polson et al. (2012) proposed an alternative Gibbs sampler for logistic and negative binomial models. The approach introduces a vector of latent variables, Z_i, that are scale mixtures of normals with independent Pólya-Gamma precision terms rather than Gamma precision terms as in the t-link model.

A random variable ω is said to have a Polya-Gamma distribution with parameters b>0 and c∈ ℜ, if

$$\omega\sim PG(b,c) \overset{m}{=} \frac{1}{2\pi^2}\sum_{k=1}^{\infty}\frac{d_k}{(k-1/2)^2+c^2/(4\pi^2)} \tag{9}$$

where $d_k's$ are independently distributed according to a Ga(b,1) distribution giving an important property of the PG(b,c) density –namely that for a∈ $\Re$ and $\eta\epsilon\Re$,

$$\frac{(e^{\eta})^a}{(1+e^{\eta})^b} = 2^{-b}e^{k\eta}\int_0^{\infty} e^{-\omega\eta^2/2}\rho(\omega/b,0)\,d\omega \tag{10}$$

Where k = a-b/2 and $\rho(\omega/b,0)$ denotes a PG (b,0) density.

Here the ridge – type estimator of β was examined for the logistic model. Different levels of ridge-type parameter (k), namely Hoerl and Kennard (1970), Lukman and Ayinde (2017) and Fayose and Ayinde (2019) was introduced, and the posterior mean was examined.

2.4. Likelihood

This is the joint probability density function (p.d.f) for the model

Logistic f $(x;\mu,s)$ $\frac{e^{-(X-\mu)/s}}{S\left(1+e^{-(X-\mu)/s}\right)^2}$ $x\in(-\infty,+\infty)$ $s>0$

A Gibbs sampler for logistic models was proposed by Polson *et al.* (2012). This involves the use of a vector of latent random variables Y_j, which are scale mixtures of normal with independent polya-gamma precision terms rather than Gamma precision terms as in t-link models. A polya-gamma random variable X with parameters (a,b) with $a>0$ and $b\in\Re$ is given as:

$$f(x|a,b) = \frac{1}{2\pi^2}\sum_{k=1}^{\infty}\frac{\omega_k}{\left(k-\frac{1}{2}\right)^2+\frac{b^2}{4\pi^2}} \tag{11}$$

Where ω_k's are independently distributed according to a $Gamma(b,1)$ distribution.

They further established a germane property of polya-gamma density that made it useful as a sampler for logistic model:

$$\frac{(e^{\eta})^a}{(1+e^{\eta})^b} = 2^{-b}e^{\kappa\eta}\int_0^{\infty} e^{-\frac{\omega\eta^2}{2}}p(\omega|b,0)d\omega, \tag{12}$$

Here, $\kappa = a-\frac{b}{2}$ and $p(\omega|b,0)$ denotes a polya-gamma density with parameters$(b,0)$.

The integrand on the right hand side is the kernel of a normal density with precision ω (i.e the conditional density of ŋ) times the prior for ω.

The LHS of equation (12) has the same function form as the probability parameter logistic regression model.

Hence, from the likelihood of binary response vector, the Bernoulli likelihood has the same form as the LHS of equation (1). So, with these properties of polya-gamma and its connection with logistic regression model, Polson et al (2012) shows that the full conditional distribution of β given Y and ω is

$$p(\beta|Y=y,\omega) \propto \pi(\beta)\exp\left[-\frac{1}{2}(z-X\beta)^T W(z-X\beta)\right] \tag{13}$$

It is clear that the random variable Z follows Normal distribution with mean ν and a variance $W^{-1}=\tau I$.

Thus, assuming a $N_p(\beta_0,T_0^{-1})$ prior for β, the full conditional for β given $Z=z$ and W is

$$N_p(m,V),$$

where $V=(T_0+X^TWX)^{-1}$, $m=V(T_0\beta_0+X^TWz)$, different K in the ridge used give rise to having the mean , $m=V(R_0+T_0\beta_0+X^TWz)$, and the variance , $V=(R_0+T_0+X^TWX)^{-1}$,

where the R_0 is the k of the different ridge parameters used.

The ridge parameters estimators used are namely

Ridge Estimator (Hoerl And Kennard, 1970a)

$$\hat{k}_i(HK) = \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2} \text{ , i = 1, 2, 3, p.}$$

Where $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p}$ and it is the Mean Square Error from the OLS regression, α_i is the i[th] element of the vector, and is also the regression coefficient from the OLS regression. $\alpha_i = Q^l \hat{\beta}$ where Q is an orthogonal matrix. p is the number of regressors and n is the sample size .

Ridge Estimator (Lukman And Ayinde, (2017)

$$\hat{k}_i(LA) = \frac{\hat{\sigma}^2}{\lambda_i \hat{\alpha}_i^2}$$

Where λ = $(\lambda_i) = 1,2,3,\ldots,p$

Ridge Estimator (Fayose And Ayinde, (2019)

$$KGRFA = \hat{k}_i^{Min}(FA) = \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}\left\{\left[\left(\frac{\hat{\alpha}_i^4 \lambda_{Min}^2}{4\hat{\sigma}^2}\right) + \left(\frac{6\hat{\alpha}_i^4 \lambda_{Min}}{\hat{\sigma}^2}\right)\right]^{\frac{1}{2}} - \left(\frac{\hat{\alpha}_i^2 \lambda_{Min}}{2\hat{\sigma}^2}\right)\right\}$$

Where $\lambda_{Min} = Min(\lambda_i) = 1,2,3,..., p$

2.5. Design

In this study, different high levels of collinearity among regressors were chosen to be:

High Positive Collinearity (HPC) when ρ = 0.80, 0.85, 0.90, 0.95,0.99 and 0.999.

Sample sizes: 10,20,30,50,100,200,300 and 500

Three explanatory variables were used for the different levels of multicollinearity with increasing sample sizes after which seven explanatory variables were also used for the different levels of multicollinearity with the increasing sample sizes.

In this study, model estimation was carried out using Bayesian approach via the Gibbs sampler of the Markov Chain Monte Carlo simulation.

3. Results and discussion

The following results were obtained for

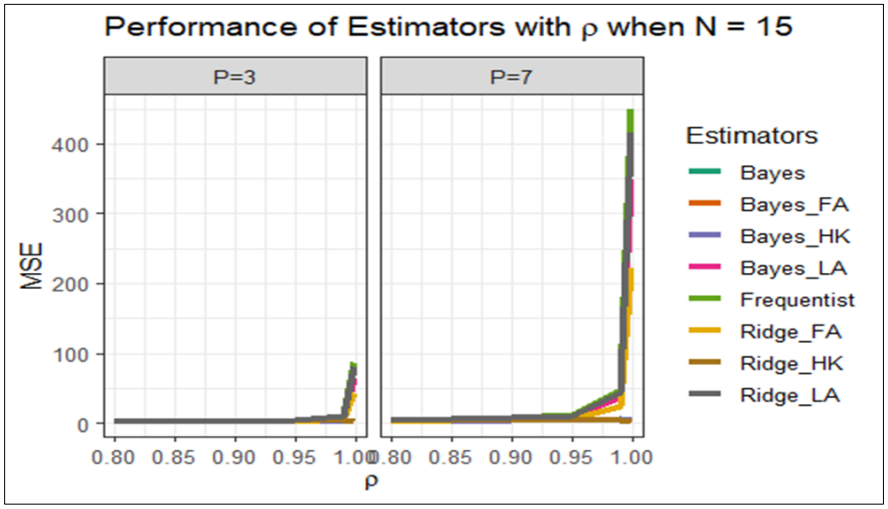

Figure 1 Performance of Estimators as Multicollinearity increases when the sample size is small

From the graph, Bayes_FA poses to perform most efficiently. Also, from the graph, the range of MSE increases with number of regressors. It can also be seen that the MSE increases with the value of ρ.

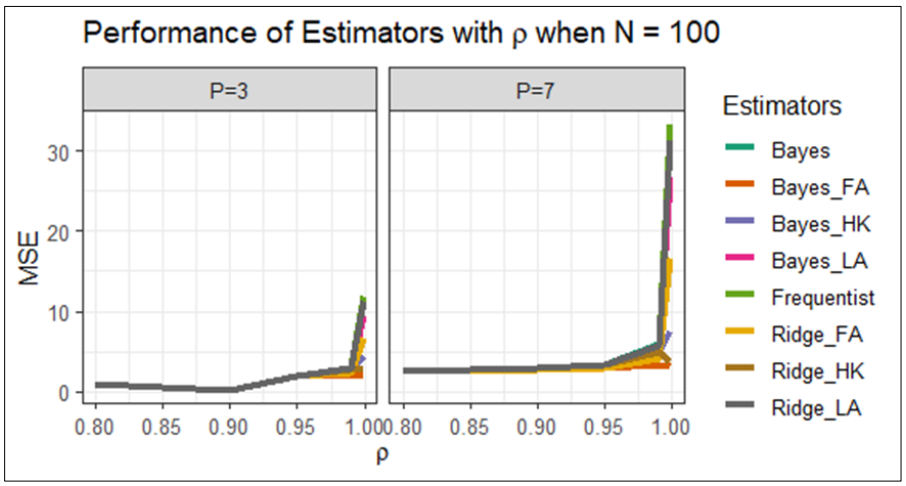

Figure 2 Performance of Estimators as Multicollinearity increases when the sample size is large

The major difference between this graph and the first graph is that the lines are more clustered when the sample size is high, this explains the reduction of variability with larger sample size. Also, the range of MSE reduces with increased sample size.

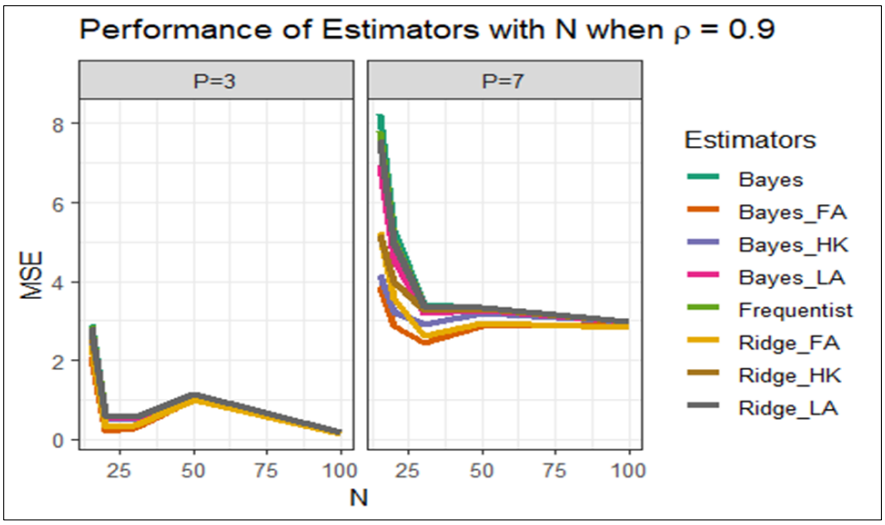

Figure 3 Performance of Estimators as Sample size increases when R=0.9

When the sample size is low ($N = 10$), the MSE is at maximum. MSE reduces towards minimum at around $N = 30$, moves up a bit and for all estimators, converges towards a point as sample size goes to infinity. This is the same for higher level of multicollinearity, but the range of MSE increases for high levels.

4. Conclusion

Focusing on the Performance of Estimators as Multicollinearity increases when the sample size is small**,** Bayes_FA poses to perform most efficiently having the range of MSE increasing with number of regressors. On the other hand the Performance of Estimators as Multicollinearity increases when the sample size is large, the graph shows that the lines are more clustered when the sample size is high, this explains the reduction of variability with larger sample size and the range of MSE reduces with increased sample size.

When the sample size is low ($N = 10$), the MSE is at maximum. MSE reduces towards minimum at around $N = 30$, moves up a bit and for all estimators, converges towards a point as sample size goes to infinity. This is the same for higher level of multicollinearity, but the range of MSE increases for high levels.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Adepoju, A.A and Ojo,O.O (2018) Bayesian method for solving the problem of multicollinearity in regression, Afrika Statistika,13:1823-1834

[2] Dawoud, I.; Kibria, B.M.G.(2020) A New Biased Estimator to Combat the Multicollinearity of the Gaussian Linear Regression Model. Stats 2020, 3, 526-541

[3] Fayose, T. S. and Ayinde, K. (2019), Different forms biasing parameter for generalized ridge regression estimator. International Journal of Computer Applications, 181(37), 21-29.

[4] Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018.

[5] Hans, C., & Van Dijk, H. K. (2010). The Bayesian Lasso for Cox Models.

[6] Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E., ... & Inger, R. (2018). "A brief introduction to mixed effects modelling and multi-model inference in ecology."

[7] Hoerl, A. and Kennard, R. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12, 55-67

[8] Park, T., & Casella, G. (2008). The Bayesian Lasso. Journal of the American Statistical Association ,June 2008, Vol. 103, No. 482, Theory and Methods doi 10.1198/016214508000000337

[9] Polson, N. G., & Scott, J. G. (2010). "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction."