A COMPARATIVE APPROACH ON BRIDGE AND ELASTIC NET REGRESSIONS

Kelachi P. Enwere[1] and Uchenna P. Ogoke[2*]

[1]Department of Mathematics and Statistics, Faculty of Science, University of Port Harcourt.

Email: kelachi_enwere@uniport.edu.ng

[2]Department of Mathematics and Statistics, Faculty of Science, University of Port Harcourt.

*Corresponding Email: uchenna.ogoke@uniport.edu.ng

Cite this article:

Enwere K.P., Ogoke U.P. (2023), A Comparative Approach on Bridge and Elastic Net Regressions. African Journal of Mathematics and Statistics Studies 6(2), 67-79. DOI: 10.52589/AJMSS-LBJO9UCU

Manuscript History

Received: 8 Jan 2023

Accepted: 16 Feb 2023

Published: 15 March 2023

ABSTRACT: *Machine Learning techniques such as Regression have been developed to investigate associations between risk factor and disease in multivariable analysis. However, multicollinearity amongst explanatory variables becomes a problem which makes interpretation more difficult and degrade the predictability of the model. This study compared Bridge and Elastic Net regressions in handling multicollinearity in multivariable analysis. Wisconsin Diagnostic Breast Cancer data was used for comparison for model fit and in handling multicollinearity between the regression techniques. Comparison were made using MSE, RMSE, R^2, VIF, AIC and BIC for efficiency. Scatter plots was employed to show fitted regression models. The results from the study show that, the Bridge regression performed better in solving the problem of multicollinearity with VIF value of 1.182296 when $\gamma = 2$ compared to Elastic Net regression with a VIF value of 1.204298 respectively. In comparison for best model fit, Bridge regression with $\gamma = 0.5$ performed better with MSE of 11.58667, AIC value of 258.9855 and BIC of 277.2217 respectively. Consequently, we can conclude that both the Bridge and Elastic Net Regressions can be used in handling multicollinearity problems that exist in multivariable regression analysis. Information on machine learning such as this, can help those in the medical fields to improve diagnosis, narrow clinical trials and biopsy to proffer effective treatment.*

KEYWORDS: Regression, Bridge, Elastic Net, Tumor Texture, Multicollinearity.

INTRODUCTION

A multivariable analysis is the most popular approach when investigating associations between risk factors and disease. In the case that a patient is diagnosed with cancer, the malignant mass must be excised. After this or a different post-operative procedure, a prediction of the expected course of the disease must be determined. Towards these considerations, machine leaning techniques have been developed targeted to provide the same levels of accuracy and prediction, without the negative aspects of surgical biopsy. One of such supervised machine learning techniques is the Regression. [1]

Hence, in multivariable analysis, regression modeling is a commonly used statistical method which allows medical researchers to examine the effects variables have on each other when multiple predictive variables are considered to estimate the association with study measurements. However, the efficiency of multivariable analysis highly depends on correlation structure among predictive variables since inference for multivariable analysis assumes that all predictive variables are uncorrelated. Thus, when the covariates in the model are not independent from one another, that is, one or more explanatory variable is determined by other variable then multicollinearity problems arise in the analysis, which leads to biased coefficient estimation and a loss of power. Numerous studies in epidemiology, genomics, medicine, marketing and management, and basic sciences have reported the effects and diagnosis of multicollinearity amongst their study variables. [2, 3].

Multicollinearity also inflates the estimates of standard errors of regression coefficients causing wider confidence intervals and increasing the chance to reject the significant test statistic. This leads to imprecise estimates of regression coefficients and false, nonsignificant p-values, and degrading the predictability of the model. [4]. Multicollinearity amongst explanatory variables can cause values of least squares estimators to be unstable, subject to change with slight variation in the data which makes interpretation more difficult since there is a lot of common variation in the variables. [5].

Though, it is not possible to eliminate multicollinearity completely but the effect of collinearity on study variables can be investigated by adopting regularized regression techniques such as bridge regression, elastic net regression, etc. [6]. Therefore, this study will explore Bridge regression and Elastic Net regression which performs best as a method for handling multicollinearity problem in multivariable regression analysis.

Consequently, the behavior of the time series variables used in regression analysis, such as nonstationary and nonlinear, and multicollinearity problem may affect the prediction accuracy in model selection. [7]. Thus, better fitted models can help increase the accuracy of predictions and improve performance of estimates of regression coefficients. Hence, it was found necessary to predict Tumor Texture of Cancerous cells from known determining factors in comparison with bridge and elastic net regression models.

Article DOI: 10.52589/AJMSS-LBJO9UCU
DOI URL: https://doi.org/10.52589/AJMSS-LBJO9UCU

LITERATURE

Recent studies have focused on using the Elastic Net as well as the Bridge method combined with other established statistical regression or forecasting methods to model the relationship between independent and dependent variables. The choice of regression algorithm to be used should be selected based on the purpose of the analysis [8], the type of data being considered, the distribution of the data and the parameters under considerations. [1] This method has been successfully applied in several scientific fields such as Medicine, Epidemiology, Genomics, and Machine learning [1 - 3]

However, multicollinearity is a serious problem that should be resolved before starting the process of data modeling. Hence, all regression analysis assumption should be met as they contribute to accurate conclusions and helps to make inferences on the population. [9] The use of the correlation coefficients and the variance inflation factor, and the eigenvalue of the covariate matrix helps one to determine the presence of multicollinearity. [10]

Reference [11] carried out a study on effects of multicollinearity in the multivariable analysis. They stated that a multivariable analysis is the most popular approach when investigating associations between risk factors and disease. Regardless of the type of dependent outcomes or data measured in a model for each subject, multivariable analysis considers more than two risk factors in the analysis model as covariates.

Reference [7], stated that Elastic net regression is a hybrid statistical technique used for regularizing and selecting necessary predictor variables that have a strong effect on the response variable and deal with multicollinearity problem when it exists between the predictor variables. Elastic Net can remove or select the predictor variables that have a high correlation in the final model and enhance the prediction accuracy.

Reference [12] carried out a research work on Bridge regression. Their study shows that bridge regression adaptively selects the penalty order from data and produces flexible solutions in various settings. The numerical study shows that the proposed bridge estimators are a robust choice in various circumstances compared to other penalized regression methods such as the ridge, lasso, and elastic net and it shows superior performances in comparisons with other existing methods.

METHODOLOGY

This study compared the Bridge Regression model and Elastic Net Regression model in order to determine the best fit regression technique to produce the better performance in handling multicollinearity. The regression analysis was based on prediction on Tumor Texture from its determining factors.

For the purpose of this study, the mean values from the Wisconsin Diagnostic Breast Cancer data, was used for comparison and data analysis. Data on Tumor Texture, Lump Area, Cell Compactness, Cell Concavity, Fractal Dimension of Lump and Radius Length.

Article DOI: 10.52589/AJMSS-LBJO9UCU
DOI URL: https://doi.org/10.52589/AJMSS-LBJO9UCU

The data used for the study where tested for the presence of Multicollinearity using VIF respectively, before proceeding to apply Bridge regression and Elastic Net regression techniques to solve the problem of multicollinearity.

R software was used to perform data analysis on Bridge and Elastic Net Regressions. Comparison for best model fit between the Bridge and Elastic Net regression models was made using MSE, RMSE, R^2, AIC and BIC from the regression analysis. Scatter plots were employed to show fitted regression models on Actual values against Predicted values. The ‘glmnet’ and ‘rbridge’ R packages were employed for the study analysis.

Bridge Regression

Bridge regression estimator generalizes both ridge regression and LASSO regression estimators, because it minimizes the SSE with a λ penalty. Thus, the bridge regression method provides a way of combining parameter estimation and variable selection in a single minimization problem.

For the multiple regression model,

$$v = \delta_0 + \delta_1 w_1 + \delta_2 w_2 + \delta_3 w_3 + \delta_3 w_4 + \delta_5 w_5 + \varepsilon \quad (1)$$

Bridge regression minimizes

$$\varepsilon^2 + \lambda(\delta_1{}^{\gamma} + \delta_2{}^{\gamma} + \delta_3{}^{\gamma} + \delta_4{}^{\gamma} + \delta_5{}^{\gamma}) \qquad \lambda > 0,\ \gamma > 0 \quad (2)$$

Bridge regression model can be stated as adding a regularization term to the multiple regression model

$$v = \delta_0 + \delta_1 w_1 + \delta_2 w_2 + \delta_3 w_3 + \delta_4 w_4 + \delta_5 w_5 + \varepsilon + \lambda(\delta_1{}^{\gamma} + \delta_2{}^{\gamma} + \delta_3{}^{\gamma} + \delta_4{}^{\gamma} + \delta_5{}^{\gamma})$$
$$\lambda > 0, \qquad 0 < \gamma \leq 2, \qquad \sum_{j=1}^{r} \left|\delta_j\right|^{\gamma} \leq q \quad (3)$$

where $\lambda\sum|\delta_1 + \delta_2 + \delta_3 + \delta_4 + \delta_5|^{\gamma}$ is the regularization term. This regularization term adds a penalty to the multiple regression model which minimizes the SSE. Thus SSE is expressed as,

$$\varepsilon^2 = (v - \hat{v})^2 + \lambda(\delta_1{}^{\gamma} + \delta_2{}^{\gamma} + \delta_3{}^{\gamma} + \delta_4{}^{\gamma} + \delta_5{}^{\gamma}) \qquad \lambda > 0,\ 0 < \gamma \leq 2 \quad (4)$$

Where w_1, w_2, w_3, w_4 and w_5 are independent variables representing Lump Area, Cell Compactness, Cell Concavity, Fractal Dimension of Lump and Radius Length used to predict the response variable v represented by Tumor Texture. q is a positive parameter representing the tuning constant that controls the amount of shrinkage. γ is the shrinkage parameter.

The Bridge estimator correctly identifies zero coefficients with higher probability than the LASSO and Ridge estimators. It performs well in terms of predictive mean square errors. Bridge regression is known to possess many desirable statistical properties such as oracle, sparsity, and unbiasedness. [13]

Article DOI: 10.52589/AJMSS-LBJO9UCU
DOI URL: https://doi.org/10.52589/AJMSS-LBJO9UCU

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 6, Issue 1, 2023 (pp. 103-115)

Elastic Net Regression

Elastic Net regression is a combination of two best techniques of shrinkage regression methods, namely, Ridge regression (l_2 penalty) for dealing with high-multicollinearity problems and the LASSO regression (l_1 penalty) for feature selection of regression coefficients. [14]

The Ridge and Least Absolute Shrinkage Selection Operator (LASSO) are special cases of the Elastic Net regression.

For the multiple regression model,

$$v = \delta_0 + \delta_1 w_1 + \delta_2 w_2 + \delta_3 w_3 + \delta_3 w_4 + \delta_5 w_5 + \varepsilon \quad (5)$$

Elastic Net regression minimizes

$$\varepsilon^2 + \lambda_1(|\delta_1| + |\delta_2| + |\delta_3| + |\delta_4| + |\delta_5|) + \lambda_2({\delta_1}^2 + {\delta_2}^2 + {\delta_3}^2 + {\delta_4}^2 + {\delta_5}^2)$$
$$\lambda_1, \lambda_2 > 0 \quad (6)$$

Elastic Net regression model can be stated as adding a regularization term to the multiple regression model

$$v = \delta_0 + \delta_1 w_1 + \delta_2 w_2 + \delta_3 w_3 + \delta_4 w_4 + \delta_5 w_5 + \varepsilon^2 + \lambda_1 \sum(|\delta_1| + |\delta_2| + |\delta_3| + |\delta_4| + |\delta_5|) + \lambda_2 \sum({\delta_1}^2 + {\delta_2}^2 + {\delta_3}^2 + {\delta_4}^2 + {\delta_5}^2)$$

$$\lambda_1, \lambda_2 > 0, \quad \sum_{k=1}^{m} |\delta_k| \leq q \qquad \sum_{i=1}^{n} {\delta_i}^2 \leq p \quad (7)$$

Where $\lambda_1 \sum(|\delta_1| + |\delta_2| + |\delta_3| + |\delta_4| + |\delta_5|) + \lambda_2 \sum({\delta_1}^2 + {\delta_2}^2 + {\delta_3}^2 + {\delta_4}^2 + {\delta_5}^2)$ is the regularization term. This regularization term adds penalty terms to the multiple regression model which minimizes the SSE. Thus SSE is expressed as,

$$\varepsilon^2 = (v - \hat{v})^2 + \lambda_1 \sum(|\delta_1| + |\delta_2| + |\delta_3| + |\delta_4| + |\delta_5|) + \lambda_2 \sum({\delta_1}^2 + {\delta_2}^2 + {\delta_3}^2 + {\delta_4}^2 + {\delta_5}^2)$$
$$\lambda_1, \lambda_2 > 0 \quad (8)$$

Where q controls the amount of shrinkage for the l_1 penalty and p controls the amount of shrinkage for the l_2 penalty. l_2 penalty is used to stabilize the l_1 penalty regularization, while the l_1 penalty is used to generate a sparse model. λ_1 and λ_2 are tuning parameters which control the strength of the regularization and selection of the predictor variable.

In finance, elastic net regression have been used to define portfolios of stocks or to predict the credit ratings of corporations.

Article DOI: 10.52589/AJMSS-LBJO9UCU
DOI URL: https://doi.org/10.52589/AJMSS-LBJO9UCU

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 6, Issue 1, 2023 (pp. 103-115)
www.abjournals.org

RESULTS AND DISCUSSION

Multiple Regression Analysis Results

From the multiple regression analysis results below for tumor texture, we see that 'lump area, cell compactness and radius length' have high VIF values of 76.98, 14.15 and 93.20 respectively, indicating the presence of high multicollinearity among these independent variables. Thus, the strong correlation between those independent variables means that they can be predicted by other independent variables in the data set.

Regression Equation for Tumor Texture

Tumor Texture	=	21.4 - 0.00396 Lump Area + 18.3 Cell Compactness + 5.3 Cell Concavity - 139 Fractal Dimension of Lump + 0.48 Radius Length

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	21.4	13.0	1.65	0.102	
Lump Area	-0.00396	0.00967	-0.41	0.683	76.98
Cell Compactness	18.3	21.7	0.84	0.402	14.15
Cell Concavity	5.3	11.7	0.46	0.650	6.85
Fractal Dimension of Lump	-139	118	-1.18	0.242	7.42
Radius Length	0.48	1.02	0.47	0.637	93.20

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3.51082	17.18%	12.78%	7.05%

From the multiple analysis results above, it is known that there is presence of multicollinearity among the explanatory variables. Therefore, we then proceed to use the Bridge and Elastic Net regression analysis techniques to solve the problem of multicollinearity among the data set.

Data Analysis and Results for Tumor Texture

From figure 1, we can see that the MSE drops suddenly (regularization taken place) as lambda values decreases from 6 to 4. Also, during the regularization, the Bridge regression shrinks out three model parameters leaving two model parameters at that stage of regularization. The best value for MSE is chosen from either the vertical fitted lines. The numbers, on top of the plot indicates the number of parameters still relevant in the model at each stage of lambda.

Article DOI: 10.52589/AJMSS-LBJO9UCU
DOI URL: https://doi.org/10.52589/AJMSS-LBJO9UCU

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 6, Issue 1, 2023 (pp. 103-115)
www.abjournals.org

From figure 2, we see similar drop (regularization taken place) in MSE between lambda values of 8 to 4 for the Bridge regression when $\gamma = 2$, indicating more accurate results in our model prediction and fitting. Moreover, all the parameters were still relevant in the model.

From figure 3, regularization takes place as lambda values tends to zero. At the initial stage of regularization, only one parameter was relevant in the model. But as MSE values continue to drop toward zero, the number of relevant parameters increases.

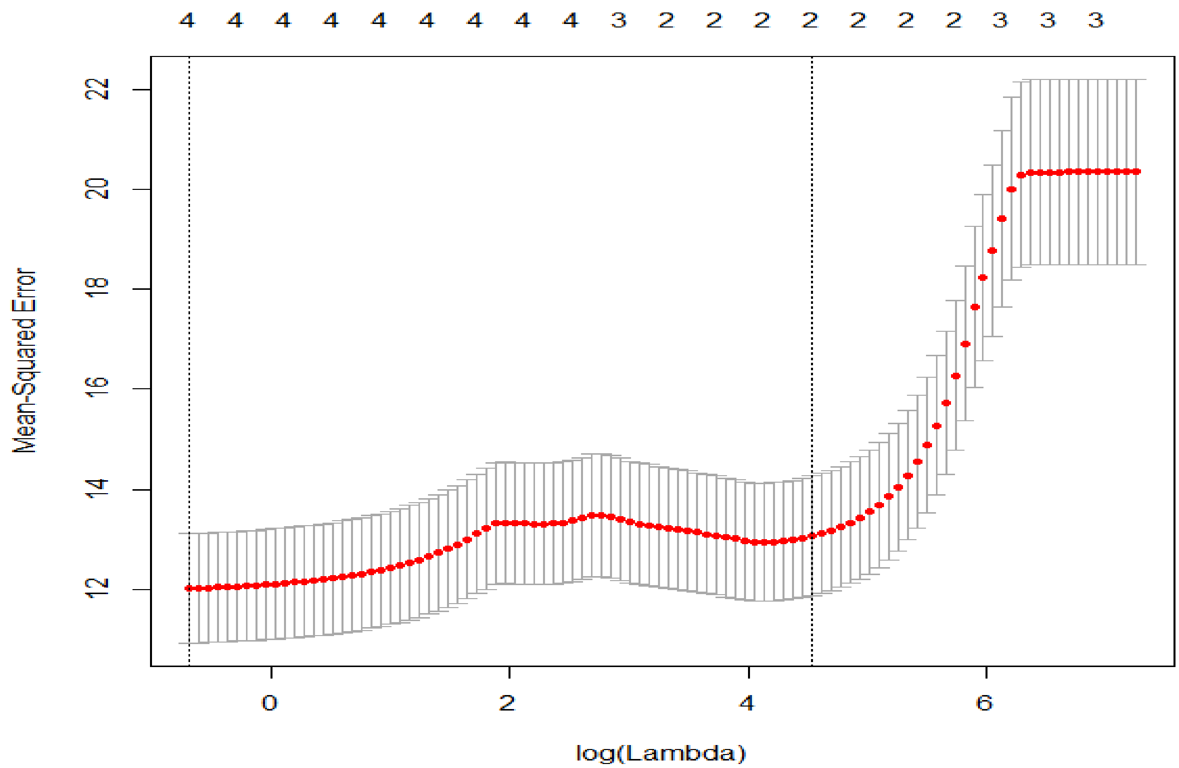

Figure 1: MSE plot for Bridge regression when γ = 1.

Article DOI: 10.52589/AJMSS-LBJO9UCU
DOI URL: https://doi.org/10.52589/AJMSS-LBJO9UCU

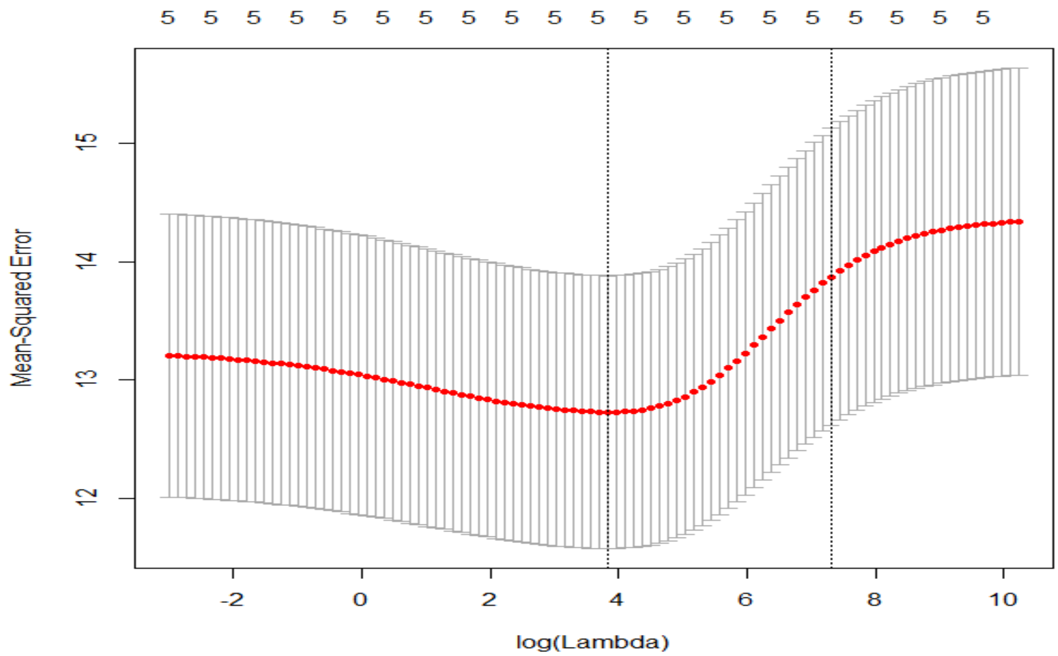

Figure 2: MSE plot for Bridge regression when $\gamma = 2$

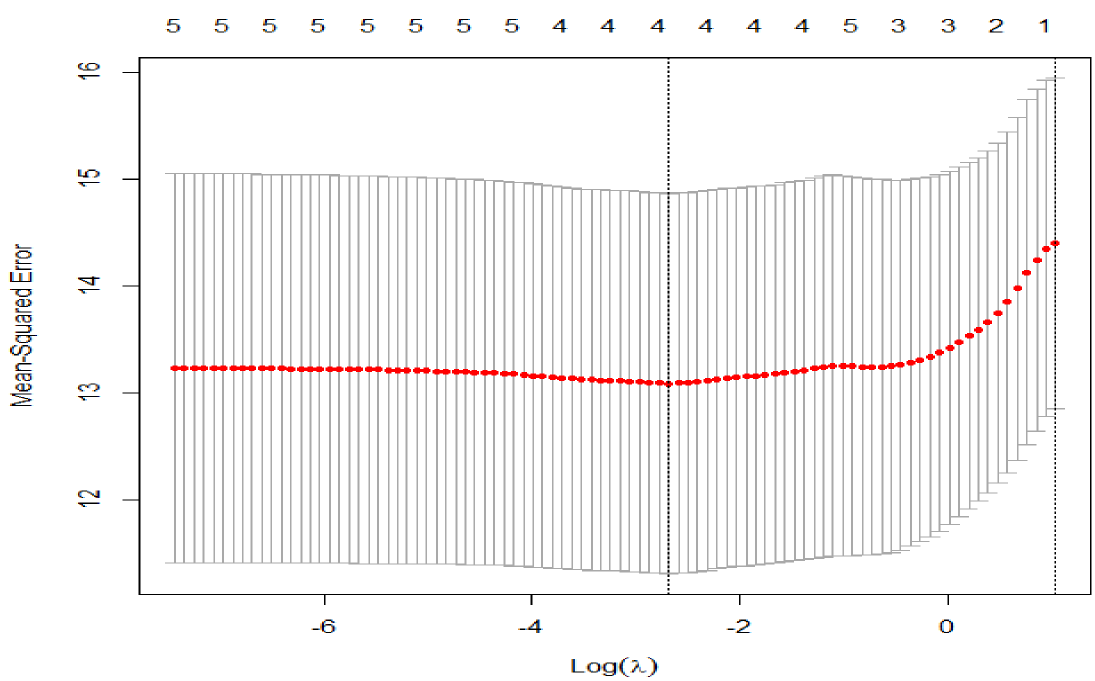

Figure 3: MSE plot for Elastic Net regression.

Article DOI: 10.52589/AJMSS-LBJO9UCU
DOI URL: https://doi.org/10.52589/AJMSS-LBJO9UCU

The Table 1 below shows the comparative analysis results of Bridge regression and Elastic Net regression for Tumor Texture, in testing for the best model fit and in handling the problem of multicollinearity.

Table 1: Comparative Results for Tumor Texture, Lump Area, Cell Compactness, Cell Concavity, Fractal Dimension of Lump and Radius Length.

Analysis Criterion	Regression Techniques				
		Bridge			
	Elastic Net	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 1.5$	$\gamma = 2$
Best λ value	0.04272611	0.5	0.5	30.67629	45.81866
MSE	11.6168	11.58667	11.58796	11.80341	11.83299
RMSE	3.408343	3.403919	3.40411	3.435609	3.439911
R^2	0.1696409	0.1717949	0.1717021	0.1563022	0.154188
$R^2{}_{adj}$	0.1254728	0.1277414	0.1276437	0.1114246	0.109198
AIC	259.2452	258.9855	258.9967	260.8389	261.0891
BIC	277.4814	277.2217	277.2329	279.075	279.3253
VIF	1.204298	1.20743	1.207295	1.185259	1.182296

Scatter Plot for Fitted Regression models on Tumor Texture

Figure 4 indicates a positive, nonlinear relationship between the actual values and the predicted values. The data points are scattered about the best fit line, indicating much variation and no correlation between the actual values and the predicted values. This data points scattered far from the best fit line are outliers. A similar representation is also seen from Figure 5.

Figure 6 indicates a positive nonlinear relationship, with no correlation between the actual values and the predicted values. Some of the data points scattered far about the line of best fit are outliers. The scatter plot for Elastic Net is somewhat similar to the Bridge when $\gamma = 2$.

Article DOI: 10.52589/AJMSS-LBJO9UCU
DOI URL: https://doi.org/10.52589/AJMSS-LBJO9UCU

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 6, Issue 1, 2023 (pp. 103-115)

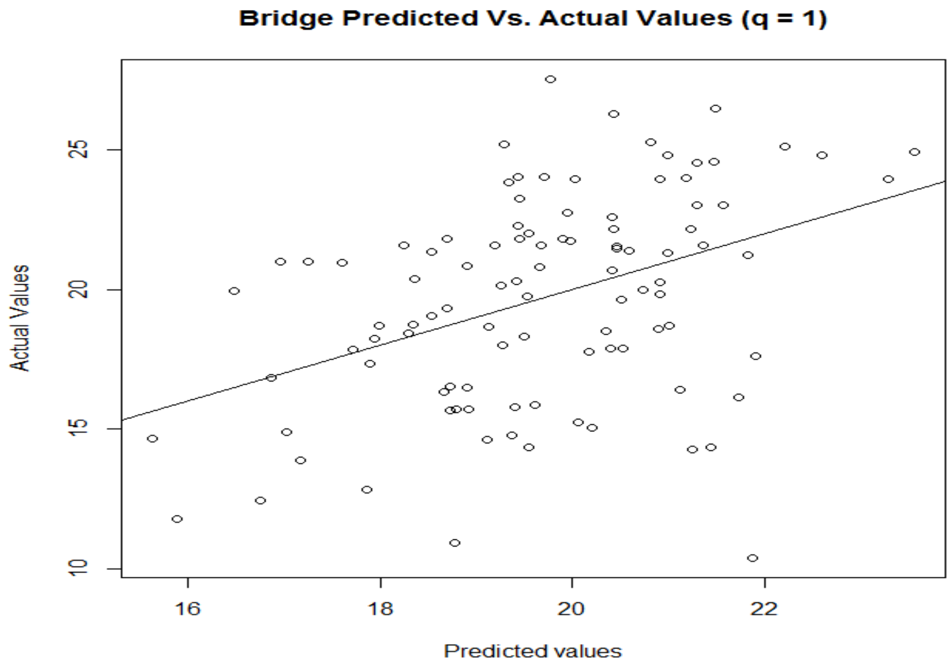

Figure 4: Bridge regression plot for Actual values against Predicted values when $\gamma = 1$

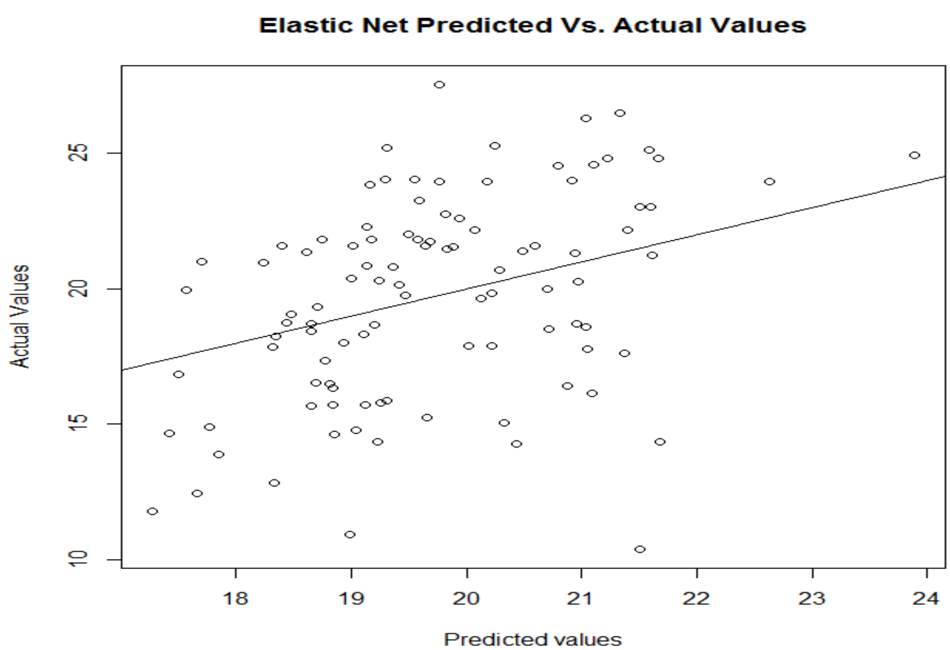

Figure 5: Bridge regression plot for Actual values against Predicted values when $\gamma = 2$

 Article DOI: 10.52589/AJMSS-LBJO9UCU
DOI URL: https://doi.org/10.52589/AJMSS-LBJO9UCU

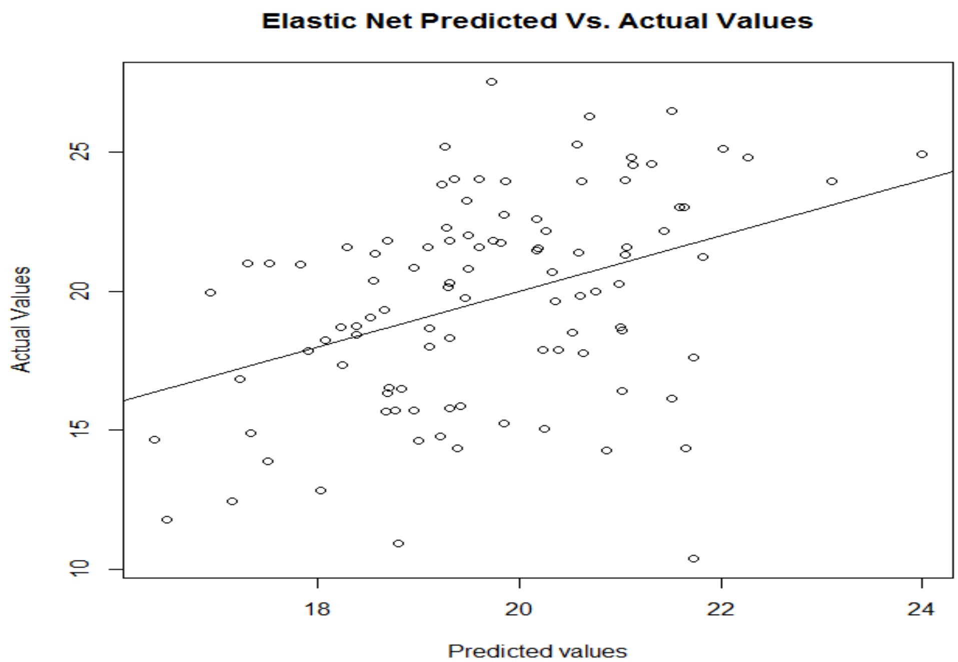

Figure 6: Elastic Net regression plot for Actual values against Predicted values

DISCUSSION

Based on our numerical results, from Table 1, when comparing the best regression technique to solve the problem of multicollinearity between Bridge and Elastic Net regressions, we see that Bridge regression performed better with VIF of 1.182296 when γ = 2 respectively. Bridge regression with γ = 2 produced better results than when γ = 0.5, 1.0, and 1.5. The Elastic Net regression also handled the problem of multicollinearity with VIF of 1.204298 respectively.

Also in our comparison for best model fit, the Bridge regression when γ = 0.5 performed better with MSE of 11.58667, AIC of 258.9855 and BIC of 277.2217 respectively. The finding is similar to [15] that the Bridge regression perform well in estimation accuracy and model selection when there are some linear restrictions present in the study. Elastic Net regression produced a MSE of 11.6168, AIC of 259.2452 and BIC of 277.4814 respectively. From the R^2 for Bridge regression when γ = 2, we have about 15% of our variation explained in the model. Thus producing more accurate results for predicting Tumor Texture using Lump Area, Cell Compactness, Cell Concavity, Fractal Dimension of Lump and Radius Length.

Article DOI: 10.52589/AJMSS-LBJO9UCU
DOI URL: https://doi.org/10.52589/AJMSS-LBJO9UCU

However, from the data set that was analyzed, the numerical results show that Bridge regression produces flexible solutions to multicollinearity in the various settings. Thus showing superior performance to Elastic Net regression which was a similar conclusion from the study done by [12].

CONCLUSION

From the study, we can conclude that both the Bridge and Elastic Net Regressions can be used in handling multicollinearity problems that exist in multivariable regression analysis. The Bridge regression with the l_1 norm is more preferred to be used in handling collinearity problems in multivariable regression, and can be used as a better technique for model fitting in order to produce better predictions.

Nevertheless, information on machine learning such as this, can help those in the medical fields to improve diagnosis, narrow clinical trials and biopsy to proffer effective treatment. The understanding of Bridge regression and the Elastic Net regression techniques will aid researchers to improve performance of estimates of regression coefficients and predict accurately possible response behaviors.

Competing Interests

The authors confirm that there is no competing of interest to declare for this publication.

REFERENCES

[1] Abhishek, T. (2021). Comparative Assessment of Regression Models Based On Model Evaluation Metrics. *International Research Journal of Engineering and Technology*, *9*(8), 853-860.

[2] Batterham, A., Tolfrey, K., & George, K. (1997). Nevill's explanation of Kleiber's 0.75 mass exponent: an artifact of collinearity problems in least squares models? *Journal of Applied Physiology*, *82*, 693-697.

[3] Parkin, I., Lydiate, D., & Trick, M. (2002). Assessing the level of collinearity between Arabidopsis thaliana and Brassica napus for A. thaliana chromosome 5. *Genome*, *45*, 356-366.

[4] Herawati, N., Nisa, K., Setiawan, E., & Nusyirwan, T. (2018). Regularized Multiple Regression Methods to Deal with Severe Multicollinearity. *International Journal of Statistics and Applications*, *8*(4), 167-172. DOI: 10.5923/j.statistics.20180804.02

[5] Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage publications.

[6] Ranjit, P. (2006). Multicollinearity: Causes, Effects and Remedies. *Indian Agricultural Statistics Research Institute*, *1*(1), 58-65.

[7] Abdullah, S., Mohd, T., & Ahmad, M. (2021). Elastic Net Regression and Empirical Mode Decomposition for Enhancing the Accuracy of the Model Selection. *International Journal of Mathematical, Engineering and Management Sciences*, *6*(2), 564-583.

Article DOI: 10.52589/AJMSS-LBJO9UCU
DOI URL: https://doi.org/10.52589/AJMSS-LBJO9UCU

[8] Serkan, E. & Mehmet, I. (2016). A Research on Comparison of Regression Models Explaining the Profitability Base on Financial Data. *International Journal of Business & Management*, *4*(10), 470-475.
[9] Jamal, D. (2017, August 8-9). Journal of Physics: Conference Series [Conference Session]. 4th International Conference on Mathematical Applications in Engineering, Kuala Lumpur, Malaysia. DOI: 10.1088/1742-6596/949/1/012009
[10] Noora, S. (2020). Detecting Multicollinearity in Regression Analysis. *American Journal of Applied Mathematics and Statistics*, *8*(2), 39-42. DOI: 10.12691/ajams-8-2-1
[11] Wonsuk, Y., Robert, M., Sejong, B., Karan, S., Qinghua, P., & James, W. (2014). A Study of Effects of Multicollinearity in the Multivariable Analysis. *International Journal of Applied Science and Technology, 4*(5), 9-19.
[12] Cheolwoo, P., & Young, Y. (2011). *Bridge regression: adaptivity and group selection.* University of Georgia. DOI: 10.1016/J.JSPI.2011.05.004
[13] Jian, H., Joel, H., & Shuangge, M. (2006). *Asymptotic Properties of Bridge Estimators in Sparse High-Dimensional Regression Models.* University of Iowa. DOI: 10.1214/009053607000000875
[14] Wang, Y., Kong, L., Jiang, B., Zhou, X., Yu, S., Zhang, L., & Heo, G. (2019). Wavelet based LASSO in Functional Linear Quantile Regression. *Journal of Statistical Computation and Simulation*, *89*(6), 1111-1130.
[15] Bahadir, Y., Mohammad, A., & Fikri, A. (2019). Penalized Regression via the Restricted Bridge Estimator. *Soft Computing*, *25*(2), 8401-8416.

Article DOI: 10.52589/AJMSS-LBJO9UCU
DOI URL: https://doi.org/10.52589/AJMSS-LBJO9UCU