World Journal of Engineering Research and Technology
WJERT
www.wjert.org
SJIF Impact Factor: 5.924

# COMPARATIVE STUDY OF MULTICOLLINEARITY USING REGULARIZATION METHOD

**Ngozi Nzelu***

Department of Statistics, Chukwuemeka Odumegwu Ojukwu University, Anambra State.

**Article Received on 06/12/2022** **Article Revised on 26/12/2022** **Article Accepted on 16/01/2023**

***Corresponding Author**
**Ngozi Nzelu**
Department of Statistics, Chukwuemeka Odumegwu Ojukwu University, Anambra State.

## ABSTRACT

In this study, Ridge, Lasso and Elastic Net Regression were compared as a regularization method to determine the model that will be better to handle multicollinearity in a dataset, especially in the area of health. Variance inflation factors were used to dictate multicollinearity in liver patient record data set and $VIF_7$ and $VIF_8$ which correspond to Total Protiens (TP) and Albumin Ratio Albumin (ALB) respectively were highly correlated. The Statistical analysis result shows that the Elastic Net Regression performed better than Ridge and lasso Regression with minimum RMSE of 0.4311013 and highest R-square of 15.62376.

**KEYWORDS:** Ridge, Lasso, Elastic Net Regression, Multicollinearity.

## 1. INTRODUCTION

Multicollinearity in regression occurs when two or more predictor variables are highly correlated to each other, such that they do not provide unique or independent information in the regression model. It can be a structural Multicollinearity caused by creating the predictors from other predictors, such as creating the predictor $x_1$ from the predictor $x_2$ or Data-based

Multicollinearity as a result of poorly designed experiment, reliance on purely observational data, or the inability to manipulate the system on which the data are collected.

Its effect can create inaccurate estimate of the regression coefficients, inflate the standard errors of the regression coefficients and it can give false non-significant p-value and degrade the predictability of model. That is, if the degree of correlation is high enough between

variables, it can cause problems when fitting and interpreting the regression model. This can cause the coefficient estimates of the model to be unreliable and have high variance.

These can be reduced by removing one or more of the violating predictors from the regression model, collect additional data under different experimental or observational conditions or It can be reduced by fitting a model containing all p predictors using a technique that constrains or regularizes or reduces the coefficient estimates or equivalently, that shrinks the coefficient estimates towards zero. The three best-known techniques for shrinking the regression coefficients towards zero are regularization methods such as ridge regression, the lasso regression and Elastic Net regression.

The aim of this paper is to determine among the regularization methods the one that is more suitable in handling multicolinearity in the area of health especially in case of liver disease without deleting or removing variable because liver is one of the most complex organs in the human body, with many functions. These include filtering blood toxins storing energy making hormones and proteins regulating cholesterol and blood sugar so adding or removing a few observations might lead to significant changes in the coefficient estimates.[3] Therefore each observation is very important to keep.

## 2. METHODOLOGY

One way of detecting multicollinearity in the regression dataset is the use of variance inflation factors (VIF), $VIF_j$ where j = 1,…, k . In practice, when $VIF_j > 5$ or $VIF_j \geq 10$ for at

least one j, it indicates a high multicollinearity for the regression matrix X.

## 3. Ridge and Lasso Regression

Ridge and Lasso regression are both known as regularization methods because they both attempt to minimize the sum of squared residuals (RSS) along with some penalty term. Using the least squares fitting procedure estimates $\beta_0, \beta_1, \dots, \beta_p$.[5]

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad (1)$$

## 4. Ridge regression

When we use ridge regression, the coefficients of each predictor are shrunken towards zero but none of them can go completely to zero. Ridge regression coefficient estimates attempts to minimize.

$$RSS + \lambda \sum_{j=1}^{p} \beta_j{}^2 \quad (2)$$

That is,

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} \beta_j{}^2 \quad (3)$$

Where *j* ranges from 1 to *p* and $\lambda \geq 0$. is called a tuning parameter. As with least squares, ridge regression seeks coefficient estimates that fit the data well, by making the RSS small. However, the second term $\lambda \sum_{j=1}^{p} \beta_j{}^2$ is known as a shrinkage penalty, it is small when

$\beta_0, \beta_1, \ldots, \beta_p$ are close to zero, and so it has the effect of shrinking penalty the estimates of βj

towards zero. The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates. When λ = 0, the penalty term has no effect, and ridge regression will produce the same coefficient least squares estimates. One of advantage Ridge regression is rooted in the bias-variance trade-off. As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.[2]

**5. Lasso Regression**

The lasso is a relatively recent alternative to ridge regression just that models generated from the lasso are generally much easier to interpret than those produced by ridge regression. However, in the case of the lasso, it's possible that some of the coefficients could go completely to zero when the tuning parameter λ is sufficiently large. Lasso regression coefficient estimates attempts to minimize

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| \quad (4)$$

That is,

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (5)$$

Where *j* ranges from 1 to *p* and $\lambda \geq 0$. $\lambda \sum|\beta_j|$ is known as a shrinkage penalty. When $\lambda = 0$,

this penalty term has no effect and lasso regression produces the same coefficient estimates as least squares. We say that the lasso is capable of producing sparse models that is, models that involve only a subset of the variables.

**6. Elastic Net regression**

Elastic Net regression is a classification algorithm that combines the properties of ridge and lasso regression method which uses a penalty function in its L1 regularization. It is a hybrid method that blends both penalizations of the L2 and L1 regularization of lasso and ridge regression.

Elastic Net regression always aims at minimizing the following loss function

$$:\frac{\sum_{i=1}^{n}(y_1 - x_i^j\widehat{\beta})^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m}\beta_j^2 + \alpha\sum_{j=1}^{m}|\hat{\beta}_j|\right) \quad (6)$$

Elastic Net also allows us to tune the alpha parameter where alpha = 0 corresponds to Ridge regression and alpha = 1 to Lasso regression. Similarly, when alpha = 0, the penalty function reduces to the L1 (ridge) regularization, and when alpha = 1, the penalty function reduces to L2 (lasso) regularization. Therefore, we can choose an alpha value between 0 and 1 to optimize the Elastic Net and this will shrink some coefficients and set some to 0 for sparse selection.[1]

**7. Analysis Summary**

This data set used in this paper was found on the UCI Machine Learning Repository, it contains 416 liver patient records and 167 non liver patient records. The data set was collected from north east of Andhra Pradesh, India. Selector is a class label used to divide into groups (liver patient or not). This data set contains 441 male patient records and 142 female patient records. There are 10 variables, we used Age as response variable while independent variables are Gender of the patient, Total Bilirubin Direct Bilirubin (TB), Alkphos Alkaline Phosphotase (AAP), Sgpt Alamine Aminotransferase (SAA), Sgot Aspartate Aminotransferase (SGAA), Total Protiens (TP), Albumin Ratio Albumin (ALB) and Globulin Ratio (AG). First we check the the multicollnearity in the dataset using variance Inflation Factor (VIF) and Lastly, we compare our result using least squares regression model and regularization model such as Lasso regression model, ridge regression model and Elastic

Net regression model to determine which model produces the lowest test RMSE and highest R2 by using k-fold cross-validation. R software were used for all Computations.

To check the data for multicollinearity, we calculate the variance inflation factors VIFj of the dataset.

**Table 1: Variance inflation factors.**

```
##   VIF1    Vif2     Vif3     Vif4     Vif5     Vif6     Vif7     Vif8
## 1.031078 4.277077 4.519834 1.118779 2.793779 2.803307 5.499190 9.894404
##    Vif9
## 3.616461
```

The $VIF_j$ values are given in the table 1 show that some of these values are rather high, like $VIF_7$ and $VIF_8$. This indicates that there are multicollinearity in the datasets since all the $VIF_j \geq 10$.

Since we determined that regularization regression is appropriate to use, we fit the model using the optimal value for λ. To determine what value to use for lambda, we perform k-fold cross-validation and identify the lambda value that produces the lowest test mean squared error (MSE). We use the **glmnet()** function in R to fit the lasso and Ridge regression model and specify alpha=1 and 0 respectively. Then setting alpha to some value between 0 and 1 is equivalent to using an elastic net regression.[4]

**The plot of test MSE by lambda value**

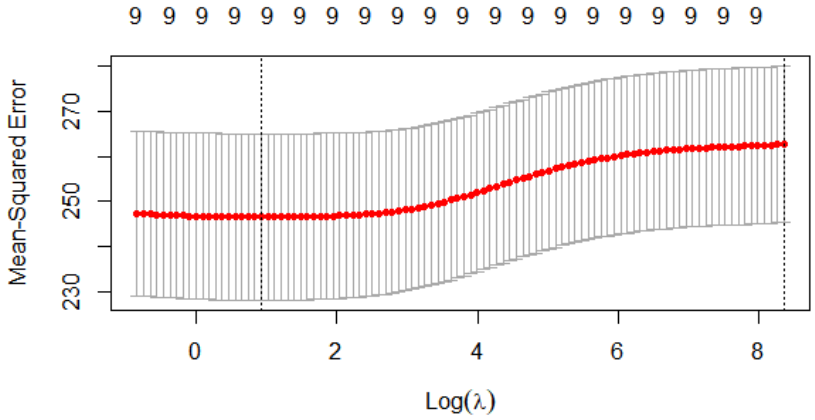


**Figure 1: Ridge Lambda value plot.**

The lambda value that minimizes the test MSE turns out to be 2.519443.

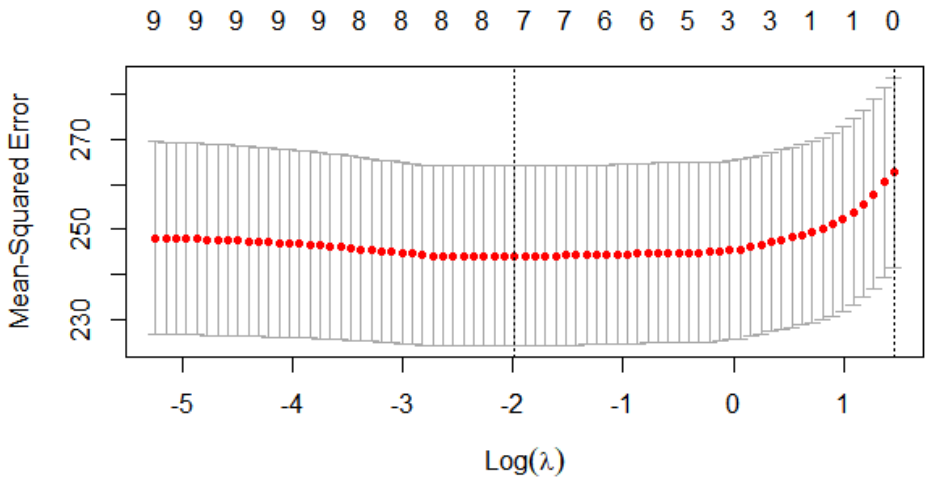


**Figure 2: Lasso Lambda value plot.**

The lambda value that minimizes the test MSE turns out to be 0.1376189.

We can analyse the model produced by the optimal lambda value by obtaining the coefficient estimates for the models.

**Table 2: The Coefficient of the model.**

```
##   b0    b1     b2     b3     b4     b5     b6    b7     b8     b9
R 61.3785 1.6136 -0.0313 -0.1623 0.0029 -0.0095 0.0021 -0.6620 -2.9723 -4.4524
L 60.3977 1.4649 -0.0004 -0.2421 0.0028 -0.0104 0.0024    0    -4.4722 -2.72016
##
```

We can see from the models that Ridge regression shrinks all coefficients towards zero while in Lasso, no coefficient is shown for the predictor Total Proteins (TP) which correspond to b7 because the lasso regression shrunk the coefficient to zero. This means it was completely dropped from the model because it wasn't influential enough.

**Elastic Net Regression**

The Elastic Net regression model is trained to find the optimum alpha and lambda values. And RMSE was used to select the optimal model using the smallest value. The final values used for the model were alpha = 0.6210763 and lambda = 0.4311013.

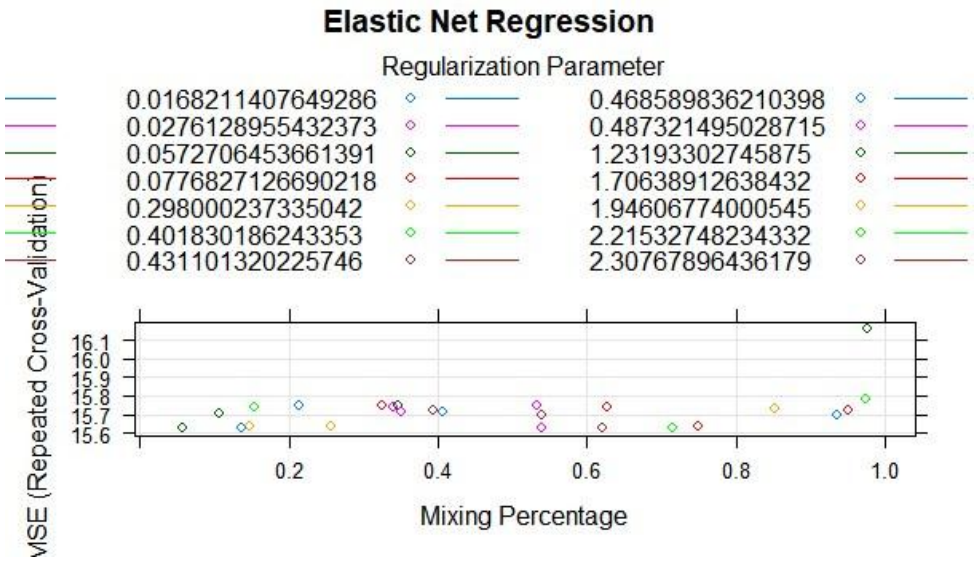


**Figure 3: Elastic Net Regression Lambda value plot.**

The mixing percentage is plotted with RMSE scores with different values of the regularization parameter and optimal value for lambda turnout to be 0.4311013.

Let us now compare the RS S values for all the models using RMSE and R2. We will first calculate the RS S on the train set and then move to the test set. 70% dataset were used for training set while 30% were used for test set and the prediction was made on the test set, the result show that.

**Table 3: The Liver Patient record Result.**

| ## | Ridge | Lasso | Elastic |
|---|---|---|---|
| RMSE | 316.9092 | 53.85248 | 0.4311013 |
| $R^2$ | -232.0504 | -5.729636 | 15.62376 |
| ## | | | |

In conclusion, the statistical analysis of Liver patients dataset show that Elastic Net Regression performed better than Ridge and lasso Regression. Therefore we can say that Elastic Net regression is a classification algorithm that overcomes the limitations of the lasso (least absolute shrinkage and selection operator) and Ridge method and its applications can be apply in other sectors of industry.

**REFERENCE**

1. Elastic Net Regression in R Programming. Jul, 28, 2020. https://www.geeksforgeeks.org

2. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning with Applications in R. Springer Texts in Statistics.
3. L.E. Melkumovaa, , S.Ya. Shatskikhb, 2017. Comparing Ridge and LASSO estimators for data analysis. 3rd International Conference "Information Technology and Nanotechnology, ITNT-2017, 25-27 April, 2017. Samara, Russia.
4. Linear, Lasso and Ridge Regression with R. Nov 12, 2019. www.pluralsight.com
5. Zach Introdution to Lasso Regression. Nov 12, 2020. https://www.statology.org