COMPARATIVE ANALYSIS OF SHRINKAGE METHODS USING SELECTED MACROECONOMIC VARIABLES IN NIGERIA

Olugbenga Ojo[1*] & Nurudeen Alabi[2]

[1,2] Department of Mathematics & Statistic s, Federal Polytechnic, Ilaro, Nigeria. P.M.B 50*
**Corresponding author's e-mail: gabriel.ojo@federalpolyilaro.edu.ng, 08037458377*

Abstract

This research presents a comparative analysis of the two most widely used shrinkage methods Ridge and Lasso Regression with Leave-One-Out Cross Validation (LOOCV) used in the determination of tuning parameters. The adoption of Mean Square Error (MSE) as a statistical loss function, where the exploration aims to single-out the optimal regression model in the face of multi-collinearity challenges was ensured. These regression model estimation techniques were compared using data on selected macroeconomic variables such as Gross Domestic Product, Transportation and Storage, Information and Communication, Construction, Trade, and Oil Refining in Nigeria. Empirically, the RIDGE and RIDGE_LOOCV techniques exhibited a superior efficiency with MSE values (0.0174 and 0.0054, respectively) compared to LASSO and LASSO_LOOCV with MSE values (0.1081 and 0.0436, respectively) afterwards, showing their enhanced performance states. Thus, the RIDGE and RIDGE_LOOCV regression techniques effectively optimized the model estimation process within the study framework. While in overall, they deduced as techniques that can best address the multi-collinearity issues within the applicability of macroeconomics.

KEYWORDS: Ridge Regression, Lasso Regression, LOOCV, Mean Square Error, Macroeconomic.

Introduction

Over the last decade, Nigeria has experienced significant improvement in macroeconomic development, leading to a boost in standards of living, and broadened opportunities for its citizens over the whole lifetimes. Moreover, trade, a mixed economy, services, communications, technology, and other indicators have acquired notable validation since raised a higher consideration as the major key to extreme growth and development because of the basic belief that most of the enlisted economic factors contributes immensely and create jobs, increase markets revenues as well as growing incomes, facilitate competition and disseminate knowledge (Ojo and Ojenike, 2020). Literarily, Eyiah-bediako et al. (2020) adopted Principal Component Analysis and Multiple Linear Regression in modeling the relationship of macroeconomic variables. In their research, it was deduced that most of the economic variables were significantly related. Arshed and Mowafaq (2021), investigated the estimation and relationship between stock market index and macroeconomic variables using traditional Ordinary Least Square and Ridge Regression. In their research, it was concluded using the mean square error criterion that ridge regression model estimated more efficiently and produced a reliable results and as well reduced the effect of multi-collinearity of the data set within the scope of the study.

Bager et al., (2017), employed the Ridge regression modeling technique in addressing the issue of multi-collinearity and to confirm the directional effect of t h e unemployment rate on macroeconomic variables. Therefore, concluded that, the study technique effectively mitigated multi-collinearity and the variables seemed significant. Ogoke and Nduka (2023), explored the Regularization methods such as Lasso and

Bridge regression to address the severe issue of multi-collinearity with the use of Corruption Perception Index data sets and Its Correlates. In this findings, efficacy of Lasso regression technique is affirmed criteria on MSE, R square, AIC and BIC, while Bridge regression performed better in terms of uncorrelated output by VIF values, showing multifaceted relationships between the variables with the study context. Brian et al., (2023), adopted the lasso and elastic net regression techniques in determining validity of the models with variables that exclusively explained poverty cases in the Economic Community of West African State (ECOWAS) countries. Hence, the study highlighted employment, government programs, fuel program, and the freedom index variables that affect concurrently the economic growth trends within the study scope.

Shady (2023), centered his research on comparison and adoption of Ridge, Elastic Net, and Lasso Regression methods in performance evaluation addressing the levels of multi-collinearity amidst multiple regression analysis of variables through simulation data sets. However, concluded that the Elastic Net method outperforms Ridge and Lasso methods in estimating the regression coefficients within the low, moderate and high level of multi-collinearity for any sample size. While, Lasso method affirms the accuracy for severe multi-collinearity with sample sizes observed below the third specification. Kelachi et al. (2023), investigated the optimal regression technique in challenging the issue multi-collinearity, evaluating the Ridge, LASSO and Bridge regression models through comparative estimation. It was concluded in their study that Bridge regression outperformed other regularized adopted techniques across the entire datasets in terms of MSE, AIC and BIC criteria respectively. In order words, a supervised statistical learning and estimation using macroeconomic data set is employed in this study to identify optimal techniques that best fit the process with a pivotal statistical loss function.

Methodology

In this study four regularized regression techniques (popularly known as shrinkage methods) such as Least Absolute Shrinkage and Selection Operator (LASSO) Regression, Ridge Regression, LASSO Leave One Out Cross Validation (LASSO_LOOCV) and RIDGE Leave One Out Cross Validation (RIDGE_LOOCV) models were compared to determine the best fit in terms of model performance (using MSE). Data on six (6) macroeconomic variables such as GDP, TRS, IFC, CON, TRA, OLR were retrieved from Central Bank of Nigeria Bulletin (1996-2021). GDP represents "Gross Domestic Product", TRS represents "Transportation and Storage", IFC represents "Information and Communication", CON represents "Construction", TRA represents "Trade", and OLR represents "Oil Refining".

Model Identification with Standardized Ordinary Least Square Approach

According to Dirk (2014), Linear regression is a learning model used for prediction and for explaining the relationship between a dependent variable (y) and an independent variable $(x_1, \dots, x_{k-1})$. However, Ordinary Least Squares (OLS) is whereby deduced as the most commonly used method for fitting the linear regression model such as;

$$y = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_{k-1} x_{k-1} + \varepsilon \quad (1)$$

where ε is specified as a random observational error.

$$y = \alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_{k-1} x_{ik-1} + \varepsilon_i, i = \overline{1, n} \quad (2)$$

Since, the measurements are expressed n times, whereby one has n values of y for n sets of x_j.

Hence, x_{ij} is the i^{th} observation for x_j, and ε_i are never directly observed.

Equation (2) after adding the parameters is transformed and expressed as a matrix form in equation (3), (4) and (5);

$$x_{10} = x_{20} = \cdots = x_{n0} = 1 \quad (3)$$

Then,

$$\boldsymbol{Y} = \boldsymbol{\alpha X} + \varepsilon, \tag{4}$$

Where,

$$\boldsymbol{Y} = [y_i]_n, \quad \boldsymbol{X} = [X_{ij}]_{n\times k}, \quad \boldsymbol{\alpha} = [a_j]_k, \quad \varepsilon = [\varepsilon_i]_n \tag{5}$$

Recall, that the coordinates $a_0, a_1, \dots, a_{k-1}$ of the vector $\boldsymbol{\alpha}$ are unknown. Then, estimating the vector $\boldsymbol{\alpha}$ as a regression model based on multivariate observations is expressed and by applying the OLS estimator in equation (5) then becomes;

$$[X, Y] = \begin{bmatrix} x_{10} & \cdots & x_{1k-1}y_1 \\ \vdots & \ddots & \vdots \\ x_{n0} & \cdots & x_{nk-1}y_n \end{bmatrix} = \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{k-1} a_j\, x_{ij}\right)^2 \mapsto min \tag{6}$$

That is, the OLS estimates of the unknown coefficients $a_0, a_1, \dots, a_{k-1}$ in equation (6) is minimized, and deduced as;

$$\hat{\alpha} = [\hat{a}_j]_k, \text{ and } det.\boldsymbol{X^T X} > 0 \tag{7}$$

Then, OLS estimates is however calculated and expressed as,

$$\hat{\alpha} = (X^T X)^{-1} X^T Y \tag{8}$$

By denotation,

$$\hat{Y} = \hat{\alpha} X \tag{9}$$

Re-writing equation (2) becomes;

$$\hat{y}_i = \hat{a}_0 + \hat{a}_1 x_{i1} + \cdots + \hat{a}_{k-1} x_{ik-1} + \varepsilon_i, \; i = \overline{1, n} \tag{10}$$

Here $\hat{y}_i$ is the predicted response value that corresponds to the predictor values $x_1, \dots, x_{k-1}$.The residual sum of squares(rss) measures the discrepancy between the data and the estimation model.

$$rss = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{11}$$

Therefore, the coefficient of determination is computed as;

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \in [0,1] \tag{12}$$

Equation (12) is used in measuring the quality of the regression model

However, applying data standardization in equation (1), and by denotation of the followings such as;

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad S_y^2 = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2, \quad S_j^2 = \sum_{i=1}^{n}\left(x_{ij} - \bar{x}_i\right)^2, j = \overline{1, k-1} \tag{13}$$

Rewritten (6) through the process of transforming using centered and normalized variables, the initial sample is modified to:

$$v_i = \frac{y_i - \bar{y}}{S_y}, \; w_{ij} = \frac{x_{ij} - \bar{x}_j}{S_y}, \quad i = \overline{1, n}, j = \overline{1, k-1} \tag{14}$$

Thus,

If $\boldsymbol{V} = [v_i]_n$, $\boldsymbol{W} = [w_{ij}]_{n\times(k-1)}$, and (15)

Since, $det\, \boldsymbol{W^T W} > 0$, then the OLS estimates for the standardized model can be denoted as:

$$\hat{\alpha} = (\boldsymbol{W^T W})^{-1}\boldsymbol{W^T V} \quad (16)$$

Model Specification and Estimation Techniques

Ridge regression Modeling Equations

The Ridge estimation for an unknown vector α in the context of standardized observations.

$\{W, V\}$ is expressed as;

$$\tilde{\alpha}_\lambda = (W^T W + \lambda I)^{-1} W^T V \quad (17)$$

$$\tilde{\alpha}_\lambda = [\tilde{a}_j(\lambda)]_{k-1} \quad (18)$$

Where I is a unity matrix, $\lambda > 0$ is expressed as the tuning ***regularization parameter***, and the Ridge estimate, indicated in coordinate form as:

$$\|V - W\alpha\|^2 + \lambda\|\alpha\|^2 \longmapsto min, \quad (19)$$

and

For all $\lambda > 0$, there is $t(\lambda) > 0$ such as that;

$$\|\boldsymbol{V} - \boldsymbol{W}\alpha\|_2 \longmapsto min \text{ subject to } \|\alpha\|_2 \leq t(\lambda) \quad (20)$$

Where, λ is the regularized parameter, $\|\alpha\|^2$ denotes the regularization function.

However, adding the regularized parameter λ to the elements of the diagonal matrix $W^T W$, then it is transformed into well-conditioned matrix $W^T W + \lambda I$. Hence, It is then demonstrated where the estimated Ridge ($\tilde{\alpha}_\lambda$) defined equivalently within the minimization problems in equation (19) and (20). Therefore, the Ridge estimate can be interpreted as an Ordinary Least Squares (OLS) estimate, but with an extra penalty introduced to the coefficient vector.

Lasso Regression Modeling Equations

Deduced from equation (17) and (18) is the estimate $\tilde{\alpha}_\lambda$, illustrating the corresponding minimization problems for standardized observations $\{W, V\}$.

Where I is the identity matrix. $\lambda > 0$ is expressed as the ***regularization parameter***, and coordinate form of the lasso estimate denoted by;

$$\|V - W\alpha\|^2 + \lambda\|\alpha\|^1 \longmapsto min, \quad (21)$$

and

For all $\lambda > 0$, there is $t(\lambda) > 0$ such as that;

$$\|\boldsymbol{V} - \boldsymbol{W}\alpha\|^2 \longmapsto min \text{ subject to } \|\alpha\|_1 \leq t(\lambda) \quad (22)$$

$$\|\alpha\|_1 = \sum_{j=1}^{k-1}|a_j|,\ a_j, j = \overline{1, k-1} \quad (23)$$

Where; a_j denotes the penalty of the coefficient vector

Ridge and Lasso Cross Validation

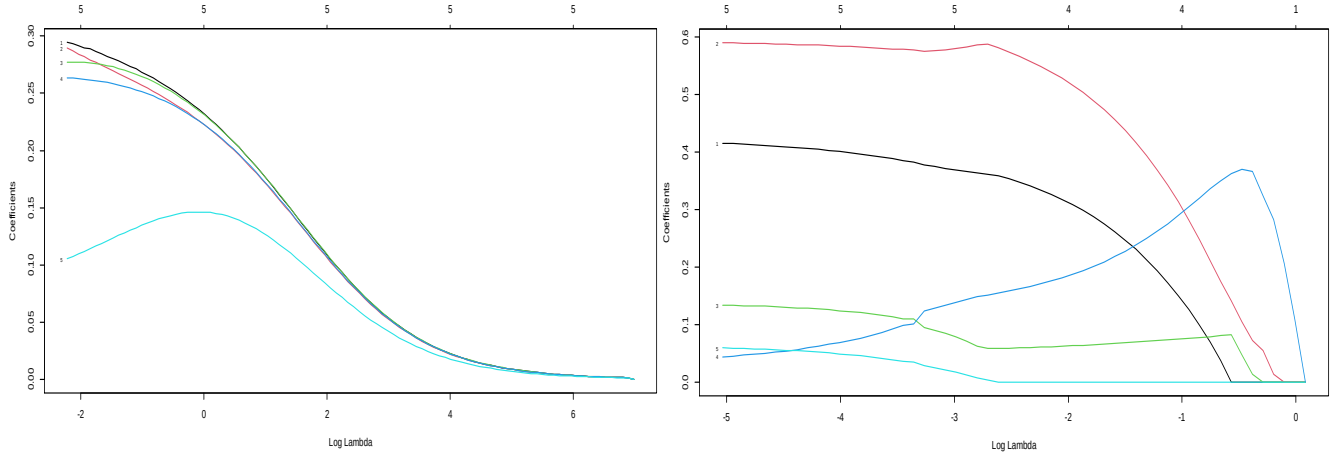
By definition, Dalip (2019) expressed cross-validation as a technique that can be employed to any supervised statistical learning method in order to deduced the method which has the least test error, low biasedness and very cheap in computations. In this study, a Leave one-out cross validation (LOOCV) is adopted where the initial data is Split into two subsets referred to as the training set and the test set.**,** estimating averagely n test creating the overall test using the means square error (mse).

Recall in equation (11) that $rss = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$, where, $\hat{y}_i = \sum_{j=0}^{k-1} \hat{b}_j(q, \lambda_s)x_{ij}$

By substitution, equation (11) becomes;

$$rss_{\lambda_s}^{q} = \sum_{i=1}^{n}(y_i - \sum_{j=0}^{k-1} \hat{b}_j(q, \lambda_s)x_{ij})^2 \quad (24)$$

where $q = \overline{1, Q}$ showed the test set as index of the block selection process.



Thence, obtaining the average rss values over all blocks is well expressed as;

$$mse_{\lambda_s} = \frac{1}{Q}\sum_{q=1}^{Q} rss_{\lambda_s}^{q} \quad (25)$$

$$cv_{(n)} = \frac{1}{n}\sum_{\lambda_s=1}^{n} mse_{\lambda_s} \quad (26)$$

Where, λ is been represented as λ_s minimizing MSE_{λ_s} estimates.

3. RESULTS

Figure 1. Ridge Trace Plot showing the macroeconomic data against $\log \lambda$

Figure 2. Lasso Trace Plot showing the macroeconomic data against $\log \lambda$

In both Ridge Regression and LASSO, the choice of the regularization or tuning parameter is crucial. If λ is too small, the regularization effect is negligible, and the model may overfit. If λ is too large, the model becomes overly constrained and may underfit. The optimal value of λ depends on the dataset and is usually determined using techniques like cross-validation or grid search. The paper employed the Leave-one-out cross validation technique in the determination of the optimal values for the tuning parameters in both models. In figure 1, the plot describes the different values of the regularization strength (λ) and as well shows the coefficients of the predictor variables (macroeconomic variables). However, as the lambda moves towards right, the coefficients tend to shrink towards zero and the absolute values of the coefficients increase. While in figure 2, this plot shows the variable selection process of LASSO regression. Thence, as the lambda increases, coefficients such as the lines labelled 4 and 5 which corresponds to the predictor macroeconomic variables; TRS and OLR shrunk to zero respectively. As features like IFC, CON, and TRA deduced with a non-zero coefficients are considered the best economic variable predictors.

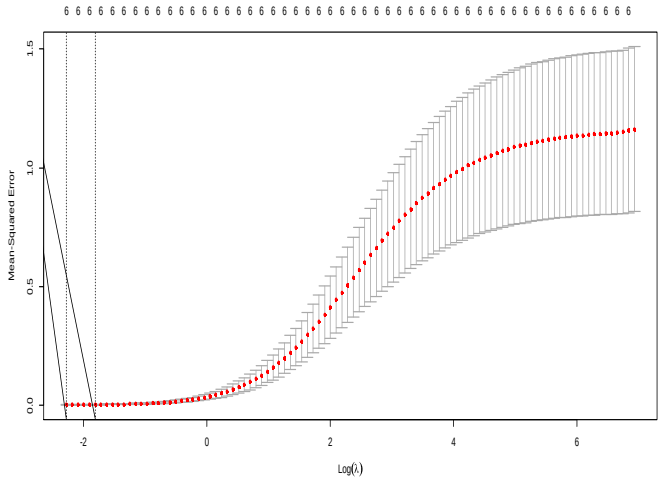


Figure 3: Depicted a RIDGE Regression's Cross-validated mean square prediction error against lambda.

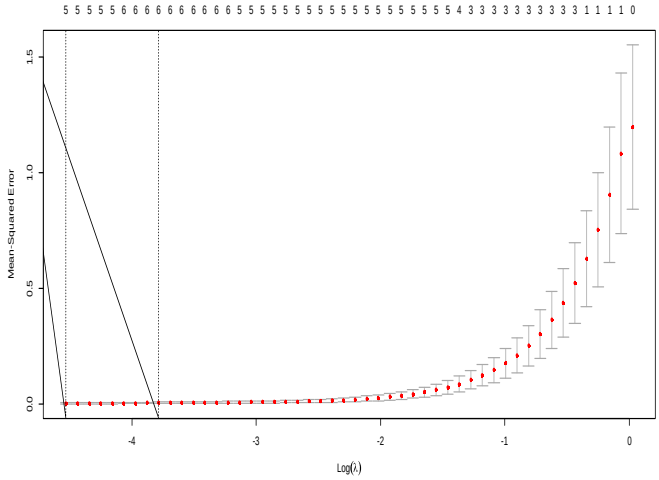


Figure 4: Depicted a LASSO Regression's Cross-validated mean square prediction error against lambda.

Figure 3 and 4, both shows the plots displaying the relationship between the mean squared prediction error (MSE) and different values of biased parameter (λ). However, the gray bars around each point deduced the MSE which represents one standard error forming addition and subtraction process. In this same vain, the left vertical dotted lines signify the minimum of MSE (λ_{min}), while the right vertical dotted lines indicates the one standard error rule from minimum (λ_{1se}). Therefore, Ridge regression gives ridge_loocv λ_{min}= 0.1029, ridge_loocv λ_{1se} = 0.1638. The LASSO yields lasso_loocv λ_{min} = 0.0108 and lasso_loocv λ_{1se} = 0.0227 correspondingly.

Table 1: Variability and Reliability of Model Techniques

Techniques for Model Estimation	Parameter(ln)	Coefficient	Standardized Error (S.E)
$RIDGE\_loocv_{\lambda_{min}=0.1029}$	**Intercept**	10.3953	**0.0009**
	TRS	0.2946	
	IFC	0.2897	
	CON	0.2772	
	TRA	0.2636	
	OLR	0.1053	
$LASSO\_loocv_{\lambda_{min}=0.0108}$	**Intercept**	10.3992	0.0023
	TRS	0.4095	
	IFC	0.5873	
	CON	0.1305	
	TRA	0.0527	
	OLR	0.0557	

Author's Source: RStudio

Table 1, shows the standard error of coefficients obtained from Ridge and Lasso regression employing Leave-One-Out Cross- Validation (LOOCV), which essentially expressed the stability and reliability of the estimated coefficients. The empirical result findings suggest, that Ridge_loocv ($\lambda_{min} = 0.1029$) with a Standard error (S.E = 0.0009) and LASSO_loocv($\lambda_{min} = 0.0108$) with a Standard error (S.E = 0.0023), wherein Ridge_loocv boasts the least standard error compared to the Lasso_loocv, indicating heightened reliability and stability in the coefficient estimates of Ridge_loocv regression (0.2946, 0.2897, 0.2772, 0.2636, 0.1053). Consequently, these outcomes affirm the significant compliance of macroeconomic variables (GDP, TRS, IFC, CON, TRA, OLR) with the precision model of RIDGE_Regression depicted in equation (27). Hence, it is concluded that ridge regression method reduces and could efficiently remove the multi-collinearity problem between macroeconomic variables within scope of the study.

$$ln(GDP) = 10.3953 + 0.2946ln(TRS) + 0.2897ln(IFC) + 0.2772ln(CON) + 0.2636ln(TRA) + 0.1053ln(OLR) \quad (27)$$

Table 2: Model Estimation Performance Summary

Estimation Techniques	MSE
RIDGE	**0.0174**
LASSO	0.1081
RIDGE_loocv	**0.0054**
LASSO_loocv	0.0436

Author's Source: RStudio

The results in Table 2 indicates the comparative analysis of Ridge Regression, LASSO Regression, Ridge and LASSO Regression with Leave- One-Out Cross Validation (LOOCV), adopting the Mean Square Error (MSE) as a

pivotal statistical loss function to strategically trace the optimal regression model on the face of multi-collinearity challenges. Thence, it can be seen that the two adopted techniques provide significant results showing that all macroeconomic variables affect the growth economically. Furthermore, the result of the RIDGE and RIDGE_LOOCV method having lower or better value of loss function MSE value (0.0174 and 0.0054), efficiently outperform the LASSO and LASSO_LOOCV counterparts (MSE of 0.1081 and 0.0436). Thus, the RIDGE and RIDGE_LOOCV method are able to estimate the model optimally within the framework.

Discussion

Based on the empirical results, Figure 1 shows an increasing regularization strength (λ) which causes coefficients to shrink towards zero and their absolute values to rise. While Figure 2, illustrates LASSO regression's variable selection, where higher lambda leads to some coefficients such as TRS and OLR shrinking to zero, favoring variables like IFC, CON, and TRA. However, Figures 3 and 4 displayed a MSE's relation to different lambda values, with gray bars showing one standard error range. Vertical lines indicate λ_min and λ_1se. Ridge regression yields λ_min=0.1029 and λ_1se=0.1638. LASSO gives λ_min=0.0108 and λ_1se=0.0227. Table 1 deduced coefficients' stability adopting LOOCV. Ridge S.E is 0.0009, LASSO S.E is 0.0023, implying Ridge regression greater stability. This research supports macroeconomic variables in compliance with Ridge as an optimal regression model against its other counterparts, showcased in Table 2 whereby comparing Ridge, LASSO, Ridge LOOCV, and LASSO LOOCV, with RIDGE and RIDGE_LOOCV showing lower MSE (0.0174 and 0.0054) than LASSO and LASSO_LOOCV (0.1081 and 0.0436), indicating superior estimation within the framework.

Conclusion

In conclusion, this study effectively demonstrated the effectiveness of various regression techniques such as RIDGE regression, LASSO regression, and the combined LASSO and RIDGE approach enhanced by LOOCV cross-validation in handling the challenge of multi-collinearity among essential macroeconomic variables: GDP, TRS, IFC, CON, TRA, and OLR. The results unequivocally highlighted the superiority of the RIDGE and RIDGE_LOOCV methods, as indicated by their significantly lower Mean Square Error (MSE) values (0.0174 and 0.0054), outperforming the LASSO and LASSO_LOOCV alternatives (with MSE values of 0.1081 and 0.0436). Thence, this investigation firmly establishes the RIDGE and RIDGE_LOOCV techniques as optimal choices for accurately modeling the relationships within this analytical framework. These methods not only effectively tackle the multi-collinearity complexities but also contribute to generating dependable insights into the interplay among the examined macroeconomic variables.

REFERENCES

Arshed, T. O. & Mowafaq, M.T.A (2021). Using Ridge Regression to Estimate Some Variables Affecting the Iraqi Stock Exchange Index. Advances and Applications in Statistics, 69(2),191-202.

Bager, A., Roman, M., Algelidh, M. & Mohammed, B (2017). Addressing Multi-collinearity in Regression Models: A Ridge Regression Application. Munich Personal RePEc Archive, MPRA Paper No. 81390, pp1-20.

Brian, W.S., Dennis, P.& Madi, E. (2023). An application of the LASSO and elastic net regression to assess poverty and economic freedom on ECOWAS countries. Mathematical Biosciences and Engineering, 20(7): 12154–12168

Dalip, K. (2019). Ridge Regression and Lasso Estimators for Data Analysis. MSU Graduate Theses. 3380. https://bearworks.missouristate.edu/theses/3380. Pp(1-35).

Dirk, J. (2014). Linear Regression for Prediction and Relationship Explanation. Journal of Statistical Modeling, 10(2), 75-88.

Eyiah-Bediako, F., Bosson-amedenu, S. & Otoo, J. (2020). Modeling Macroeconomic Variables Using Principal Component Analysis and Multiple Linear Regression: The Case of Ghana's Economy. Journal of Business and Economic Developmens, 5(1): 1-9.

Kelachi, E., Ethelbert, N. & Uchenna, O. (2023). Comparative Analysis of Ridge, Bridge and Lasso Regression Models in the Presence of Multi-collinearity. IPS Intelligentsia Multidisciplinary Journal, 3(2): 1-8.

Ogoke, U.P. & Nduka, E.C. (2023). Comparative Analysis of Lasso and Bridge Regression Using Corruption Perception Index and Its Correlates in Nigeria. Annal Biostat & Biomed Appl.5(1),2-5.

Ojo, O. and Ojenike, O. (2020). Statistical Modeling of Foreign Trade on Economic Growth and Development: A Case Study of Nigeria. Proceedings of the 2nd International Conference, the Federal Polytechnic, Ilaro. pp269-279.

Shady, I.A. (2023). Evaluation of Ridge, Elastic Net and Lasso Regression Methods in Precedence of Multicollinearity Problem: A Simulation Study. Journal of Applied Economics and Business Studies, 5(1),131-142.