OPEN ACCESS
Journal of Multidisciplinary Applied Natural Science Vol. 5 No. 2 (2025) https://doi.org/10.47352/jmans.2774-3047.251
Check for updates
Research Article

Performance of Ridge Regression, Least Absolute Shrinkage and Selection Operator, and Elastic Net in Overcoming Multicollinearity

Dewi Retno Sari Saputro*, Nugroho Lambang Wahyu, and Yekti Widyaningsih

Received : July 15, 2024	Revised : September 25, 2024	Accepted : November 19, 2024	Online : February 23, 2025

Abstract
Multicollinearity is a violation of assumptions in multiple linear regression analysis that can occur if there is a high correlation between the independent variables. Likewise, the variants of multiple linear regression models such as the Geographically Weighted Regression model (GWR). Multicollinearity causes parameter estimation using the Quadratic Method (QM) unstable and produces a large variance. On the other hand, what is expected in the estimation parameters is an estimate with a minimum variance, even though it is biased. Thus, one way to overcome multicollinearity can be to use biased estimators, such as Ridge Regression (RR), Least Absolute Shrinkage and Selection Operator (LASSO), and Elastic Net (EN). In RR, the Least Square Method (LSM) coefficient is reduced to zero but it can't select the independent variable. However, the parameter model obtained from the Ridge Regression is biased, and the variance of the resulting regression coefficients is relatively tiny. In addition, the RR is increasingly difficult to understand if a huge number of independent variables are used. Meanwhile, LASSO is a computational method that uses quadratic programming and can act out the RR principles and perform variable selection. The LASSO method became known after discovering the Least-Angle *Regression* (LARS) algorithm. The LASSO method can reduce the LSM coefficient to zero to perform variable selection. LASSO also has a weakness, so EN is used. In this article, the performance of the three methods is compared from the mathematical aspect. The performance of each is written as follows, RR is helpful for clustering effects, where collinear features can be selected together; LASSO is proper for feature selection when the dataset has features with poor predictive power and EN combines LASSO and RR, which has the potential to lead to simple and predictive models.

Keywords: ridge regression, LASSO, elastic net, multicollinearity

1. INTRODUCTION

In multiple regression analysis, the relationship between two or more independent variables is often a problem; namely, there is a mutual correlation between independent variables. It is said to be double collinearity (multicollinearity). In principle, multicollinearity means that there is a linear relationship between some or all of the independent variables of a regression model [1][2]. In practice, it is possible to find correlated independent variables [3][4]. With multicollinearity, parameter estimation using the Least Squares Method (LSM) becomes invalid, resulting in a violation of the multiple regression model assumption [5]. The fact that some or all of the independent variables are correlated with LSM does not prevent obtaining a suitable regression function or affect inferences about the mean of the dependent variable or the forecasting of new observations, as long as the inference is carried out within the observed area. High multicollinearity results in the regression parameter estimators tend to have a large diversity, meaning that the estimated regression parameters tend to vary significantly from one sample to another [6]. This resulted in not obtaining appropriate information about the actual regression parameters (population). If the collinearity is in some independent variables, it can be overcome using the stepwise regression method so that the variables that cause collinearity are excluded from the equation [7]. On the other hand, if all independent variables are highly correlated, it is not easy to do and will not give a good solution.

Multicollinearity resulted in the regression model estimator no longer being BLUE. This is because the estimator variance (β) of the regression model is large. This results in hypothesis testing regarding the significance of the regression coefficient estimator tends to accept H_0, which means that the regression coefficient estimator is

Publisher's Note:
Pandawa Institute stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Table 1. Comparison of l_1 - regularization and l_2 - regularization.

No.	l_1 - ***regularization***	l_2 - ***regularization***
1.	Penalizes the sum of the absolute value of weights.	Penalizes the sum of square weights.
2.	It gives multiple solutions.	It has only one solution.
3.	It is constructed in feature selection.	There is no feature selection.
4.	It generates simple and interpretable models.	It gives more accurate predictions when the output variable is the function of the whole input variables.
5.	Computationally inefficient over non-sparse conditions.	Computationally efficient because of having analytical solutions.

not significantly different from zero. Multicollinearity causes the diagonal elements of the $\boldsymbol{X}^T \boldsymbol{X}$ matrix to be large and almost singular (close to zero), where $\boldsymbol{X}$ is a matrix of independent variables [8]. This resulted in the instability of the regression model so that the regression coefficient could not be determined [9]. Multicollinearity also leads in the presence of a small characteristic root/ eigenvalue, causing the estimator's variance to increase. As a result of multicollinearity, multiple linear regression modeling with the Least Squares Method is rendered incorrect, despite the fact that the estimator remains unbiased. As therefore, an approach is required to overcome multicollinearity.

Several approaches can be taken to avoid violating the multicollinearity assumption. One approach to overcome multicollinearity is Ridge Regression, Least Absolute Shrinkage and Selection Operator (LASSO), and Elastic Net (EN). The principle of the three settlements is to provide penalties to improve Least Squares Method (LSM) performance. Ridge Regression is used to overcome multicollinearity in multiple linear regression models by adding ridge parameters [10][11]. Determination of ridge parameters using the Generalized Cross-Validation (GCV) approach (He et al., 2022. Another model approach is with LASSO. Tibshirani (1996) [8] states that LASSO estimates using the LARS algorithm will reduces the parameter estimate to zero. The LASSO method is used to overcome local multicollinearity cases so, it is hoped that an unbiased and efficient parameter estimate can be obtained [12].

The multicollinearity problem in regression models can be solved using several methods, each with specific advantages and disadvantages. Çelik et al. [13] states that to overcome multicollinearity in the regression model, the LASSO method is considered more consistent even though the predictor variables have a high level of multicollinearity compared to the Ridge method. Several studies state that EN improves Ridge and LASSO performance. In this study, the performance of the three methods is compared from the mathematical aspect. The study compares the performance of Ridge Regression, LASSO, and Elastic Net in overcoming multicollinearity in multiple linear regression models. It aims to demonstrate how these methods address the issues caused by multicollinearity and evaluate their effectiveness in producing stable, predictive models. The study seeks to highlight the strengths and weaknesses of each method in terms of handling correlated features, feature selection, and variance reduction, with a focus on understanding their theoretical and practical applications.

Specifically, this article is divided into numerous parts. The first section defines multicollinearity and the theory behind handling multicollinearity situations. The second section explains the materials and procedures, while the third section describes the findings and discussion. This part will go over numerous topics, including the features of each RR, LASSO, and EN method, the distinctions between the three methods, the benefits and drawbacks of the three methods, and the real-world applications of the three RR, LASSO, and EN methods. The final section provides the conclusions, which analyze the characteristics of the three techniques.

2. MATERIALS AND METHODS

The material used as the basis for this research study includes two main things, namely multiple linear regression and the Least Square Method.

2.1. Multiple Linear Regression

Çelik et al. [13], states that multiple linear regression is one of the widely used statistical techniques involving more than one independent variable X. Multiple linear regression model is

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon_i \; ; i = 1,2, \ldots, n \quad (1)$$

where Y is the dependent variable, is the model parameter, X is the independent variable, $\boldsymbol{\beta_0}$ is intercep term (intersection point of the regression line with the Y axis), $\boldsymbol{\beta_1}, \ldots, \boldsymbol{\beta_k}$ is coefficient for the independent variable, and $\boldsymbol{\varepsilon_i}$ is the remainder with normal distribution. There are several assumptions underlying the multiple linear regression analysis; namely, there is no serial correlation (autocorrelation) on the residuals, the independent variables are constant, there is no multicollinearity between the independent variables, the residuals are normally distributed (normality), and the variance of the residuals is constant (heteroscedasticity). If there is a violation of one of the assumptions then it affects the regression estimator.

2.2. Least Squares Method (LSM)

Basak and Majumdar [14], states that the Least Squares Method is a method for estimating parameters in linear regression. The principle of LSM is to minimize the sum of the residual squares from the regression model using the Lagrange multiplier. The matrix form of multiple linear regression is

$$\boldsymbol{Y} = \boldsymbol{\beta X} + \boldsymbol{\varepsilon} \quad (2)$$

so that the sum of the squares of the remainder of the model is

$$\sum_{i=1}^{n} \varepsilon_i^{\,2} \text{ atau } \boldsymbol{\varepsilon^T \varepsilon}$$

The concept is to minimize the number of residual squares by reducing it to the model parameter to obtain a multiple linear regression estimator, namely

$$\boldsymbol{\hat{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y} \quad (3)$$

The LSM produces an estimator that is the best linear unbiased estimator (BLUE). This means that the estimator is a linear function that is unbiased and has a minimum variance. This results in the model being used appropriately.

3. RESULTS AND DISCUSSIONS

As stated earlier, multicollinearity is a situation that indicates a correlation or strong relationship between two or more independent variables in a multiple regression model. If multicollinearity occurs, then a variable that is strongly correlated

Table 2. The comparison of Ridge Regression, LASSO, and Elastic Net in overcoming multicollinearity.

Methods	Solution of Multicollinearity Problem
Ridge Regression (l_2 - regularization)	1. Adds a penalty to the square of the regression coefficient.
	2. Reduces the variance of highly correlated coefficient estimates.
	3. Stabilize the model with smaller and less extreme coefficients.
LASSO (l_1 - regularization)	1. Adds a penalty to the absolute value of the regression coefficient.
	2. Select variables by setting several coefficients to zero.
	3. Reduce variance and simplify the model by eliminating unimportant variables.
Elastic Net (combination of l_1 - regularization and l_2 - regularization)	1. Combination of L_1 and L_2 penalties.
	2. Combines the advantages of Ridge and LASSO; improve model stability and carry out variable selection.
	3. Effective for data with high correlation between variables.

with other variables in the model, its predictive power is unreliable and unstable. The impact of multicollinearity is that the Partial Regression Coefficient is not measured precisely, small changes in data from sample to sample will cause drastic changes in the value of the partial regression coefficient, and changes in one variable can cause significant changes in the value of the partial regression coefficient of other variables, the Confidence Interval value is so vast that it will be very difficult to reject the null hypothesis in a study if there is multicollinearity in the study. Therefore, one way to overcome multicollinearity is to use biased estimators, namely RR, LASSO, and EN. These three methods have unique performance to overcome the multicollinearity problem. Therefore, their performance is discussed in this paper.

3.1. *Ridge Regression*

Ridge Regression is a modification of the LSM through Lagrange multiplier to overcome multicollinearity, which introduced by Hoerl and Kennard (1970) [15] to deal with the instability of the least squares estimator. Modification is done by adding the coefficient of bias constant λ on the diagonal of the $\boldsymbol{X^TX}$ matrix [8]. Ridge Regression aims to minimize the variance and standard error of the estimator so that the significance test of the regression coefficient estimator tends to reject H_0. This means that the regression estimator is significantly different from zero, or there is a significant effect between the independent variables on the dependent. In this case, the RR estimator is no longer a BLUE (Best Linear Unbiased Estimator) because of the addition of the estimator coefficient $\boldsymbol{\hat{\beta}_R}$ to be biased even though the variance is minimal.

Ridge Regression pinpoints the size of the regression coefficient at norm l_2 or specifically predicts $\hat{\beta}$ by minimizing SSE (β) with constraint

$$\sum_{j=1}^{p} \beta_j^2 \leq t \quad (4)$$

In this form, $\sum_{j=1}^{p} \beta_j^2$represents the l_2 norm or Euclidean norm of the vector β, and t is a threshold that controls the size of the coefficients. Or, if you write it in another form that is to minimize

$$\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{k=1}^{p} \beta_k x_{ik}\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \quad (5)$$

Where y_i is the observed value for the i-th sample (dependent variable); β_0 is the intercept term; β_k is the coefficient for the k-th independent variable; x_{ik} is the value of the k-th independent variable for the i-th sample; λ is the regularization parameter or penalty term.

The selection of bias coefficient λ is the most crucial thing in RR. If the selection of the bias coefficient λ is correct, the estimator value and Variance Inflation Factor (VIF) become more stable. Stability indicates that multicollinearity can be overcome. The VIF value is stable if it is relatively close to 1 and less than ten, and the bias coefficient λ lies in the interval $0 < \lambda < 1$. To determine the exact value to obtain a stable estimator, a Ridge trace is used. The ridge trace is a graph containing the results of selecting the value of λ and its effect on the regression estimator value. If the value of $\lambda = 0$, then the magnitude of the coefficient of the Ridge estimator $\boldsymbol{\hat{\beta}_R}$ will be the same as the least squares estimator $\boldsymbol{\hat{\beta}}$. On the other hand, if $\lambda = 1$, then the Ridge estimator coefficient $\boldsymbol{\hat{\beta}_R}$ will be biased with the Least Squares estimator $\boldsymbol{\hat{\beta}}$.

The solution to the RR is obtained similarly in the Least Squares Method, namely by minimizing the Sum Squares of Error (SSE).

$$SSE(\beta, \lambda) = (\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda \boldsymbol{\beta^T\beta} \quad (6)$$

Ridge Regression is one method that can be used to overcome the problem of multicollinearity through modification of the LSM [13]. The modification is carried out by adding a relatively small bias constant on the diagonal of the $\boldsymbol{X^TX}$ matrix so that the estimating coefficient of the ridge is influenced by the magnitude of the bias constant λ. Thus, the estimated parameters are obtained using the following description. Based on equation (6), it is obtained equation

$$\boldsymbol{X^Ty} = (\boldsymbol{X^TX} + \lambda \boldsymbol{I})\boldsymbol{\beta}$$

Where $\boldsymbol{X}$ is the design matrix containing the independent variable; $\boldsymbol{y}$ is the vector of the observed outcomes (dependent variable); $\boldsymbol{\beta}$ is the vector of regression coefficients to be estimated; $\boldsymbol{X^T}$ is the trapose of the matrix $\boldsymbol{X}$; $\boldsymbol{I}$ is identity matrix.

Before parameter estimation, the model indicated by multicollinearity was transformed into

Table 3. The Comparison of Ridge Regression, LASSO, and Elastic Net in Effect and Benefit of Adding Penalties to The Regression Coefficient.

Methods	Effects	Benefits
Ridge Regression	• Drives some regression coefficients to zero • Enables automatic feature selection • Unimportant variables can be eliminated from the model	• Produces simpler models with easier interpretation.
LASSO	• Reducing the magnitude of the regression coefficient • Prevent coefficients that are large • Reducing variability in the model	• Reduces overfitting by maintaining significant coefficients.
Elastic Net	• Combination of L_1 and L_2 effects pushing some coefficients to zero and adjusting the magnitudes of other coefficients. • Provides flexibility in handling multicollinearity and feature selection.	• The combination of Ridge and Lasso's strengths, helps overcome each other's weaknesses.

a standard form. The standard form of the regression model was obtained through the procedure of centring and scaling. Centering is the difference between each observation and the average of all observations for the variable. Centering is a procedure to remove the parameter from the regression model. This procedure is carried out to simplify the process of forming a new regression model. In contrast, scaling (rescaling) is an observation of the standard deviation for the variable.

The choice of the magnitude of the bias λ constant is a problem that needs to be considered. The desired bias λ constant is a bias constant that produces a relatively small bias and produces a relatively stable estimator coefficient. Several references are used to determine the magnitude of λ, including by looking at the VIF size and looking at the trend pattern of the ridge trace. The ridge trace is a pattern of the Ridge Regression estimator together with various possible values of the bias λ constant according to Gibbons and McDonald, 1984 [16]. The selected value of λ is which gives a relatively stable Ridge Regression estimator value $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$.

Ransom et al. [15] determines the value of λ by using a ridge trace which is a plot of data between $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ with several values λ in the interval (0,1) until stability is achieved in the predicted parameter. However, selecting a Ridge trace becomes a subjective procedure because it requires the researcher's decision to determine the chosen q value, [17][18]. Hoerl Kennard, and Balwin (1975) [19] suggest selecting the value of λ by using the formula

$$\lambda = \frac{p\hat{\sigma}^2}{\hat{\beta}^t\hat{\beta}}$$

with p is the number of features in the model, $\hat{\beta}$ and $\hat{\sigma}$ obtained from the Least Squares method in further research [14], proposes an iterative procedure using the value of λ in (5) as the initial value to calculate the value of and then $\hat{\beta}$ and $\hat{\sigma}$ used are obtained from the Ridge Regression method. Thus, this procedure will stop when

$$\frac{q_{j+1} - q_j}{q_j} > 20\,T^{-1,3}$$

when q_j is a quantity that is being optimized or monitored over iterations, j and $j+1$ is the number of iterations, and T is scalar quantity calculated as follows:

$$T = \frac{Trace(\boldsymbol{X}^T\boldsymbol{X})^{-1}}{p} = \frac{\sum_{j=1}^{p}\left(\frac{1}{\lambda_j}\right)}{p}$$

Trace is the trace of a matrix (the sum of the diagonal elements of that matrix) and λ_j is the eigenvalues of the matrix $\boldsymbol{X}^T\boldsymbol{X}$.

Loss Function

$$L_{Ridge}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j=1}^{m} \hat{\beta}_j^2 = ||y - X\hat{\beta}||^2 + l_1||\hat{\beta}||^2$$

If $\lambda \longrightarrow 0$ then $\hat{\beta}_{Ridge} \rightarrow \hat{\beta}_{MKT}$ and $\lambda \longrightarrow \infty$ then $\hat{\beta}_{Ridge} \rightarrow 0$

The estimator of the Ridge Regression parameter in equation (6) is biased, not equivariant [20], meaning that the estimator will have different results if the original variable is standardized with the original variable. Therefore, it is recommended to standardize the scale of the original variable so that it has an expected value of zero and a variance of one [20]. Ridge Regression is a shrinkage method or a regression. Coefficient method that can be used to solve multicollinearity problems. Although the model obtained from the Ridge Regression is biased, the resulting coefficient estimator tends to be more stable than the LSM.

$$l_1 = \lambda \sum_{j=1}^{m} \hat{\beta}_j^2$$

l_1 or LASSO regular is a regularization technique that requires us to minimize the sum of absolute values between features and target variables. The word "absolute" here means that in example we only care about the largest value between feature and target, and other smaller values within the same feature will be ignored (or removed) when training models using this model.

3.2. *Least Absolute Shrinkage and Selection Operator (LASSO)*

The LASSO method was first introduced by Friedman et al. in 1996 [20]. The LASSO coefficient estimator is obtained using quadratic programming. LASSO is one of the regression techniques for reducing the independent variables. Still according to Friedman et al. [20], LASSO shrinks the regression coefficient of the variable with a high correlation with error, to precisely zero or close to zero. LASSO is a method of Penalized Least Squares (PLS) that converts the constraints in the Ridge Regression into the form of l_1-norm which is also known as l_1- regularization. Parameter estimation on LASSO is obtained by minimizing the following equation.

$$JKS = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{k=1}^{p} \beta_k x_{ik})^2$$

JKS seems to refer to the objective function, specifically the Sum of Squared Errors (SSE) and n is the number of observations. Then, the constraint

$$\sum_{k=1}^{p} |\beta_k| \le t$$

Loss function

$$L_{Lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j=1}^{m} |\hat{\beta}_j|$$

where λ is the tuning of the parameter whose numerical size is determined by Cross-Validation (CV) and determines the Shrinkage of the LASSO coefficient, $\lambda > 0$ and m is total number of coefficients. LASSO parameter estimation cannot be obtained in closed form as in LSM, but using quadratic programming [8]. The LASSO parameter estimate is written as

$$\hat{\beta}_{Lasso} = argmin\left\{\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{k=1}^{p} \beta_k x_{ik})^2 + \lambda \sum_{k=1}^{p} |\beta_k|\right\}$$

Estimation of LASSO parameters is obtained by determining standardized limits, namely

$$s = \frac{\lambda}{\sum_{j=1}^{k} |\hat{\beta}_j^0|}$$

where $\hat{\beta}_j^0$ represents the OLS (Ordinary Least Squares) estimates of the coefficients, λ is the regularization parameter, and s is the Shrinkage parameter, which has a value from 0 to 1 with

$$t = \sum_{j=1}^{k} |\hat{\beta}_j|$$

If $\hat{\beta}_j^0$ is the least squares parameter estimate (OLS) and $t_0 = \sum_{j=1}^{k} |\hat{\beta}_j^0|$, the value of $t < t_0$ causes several parameters to be zero. In some literature

$$\lambda \sum_{j=1}^{m} |\hat{\beta}_j| = l_2$$

Although LASSO has shown success in many situations, it has some limitations. Therefore according to Tibshirani [8] and Zöngür and Buzpinar [21], in its application it is necessary to consider the following.

1. In the $p > n$ case, LASSO selects at most n variables before saturation, due to the nature of the convex optimization problem. This seems to be a limiting feature for the variable selection

method. Moreover, LASSO is not well defined unless bound to the norm l_1 of a coefficient smaller than a certain value.
2. If there is a group of variables in which the pairwise correlation is very high, then LASSO tends to choose only one variable from the group and does not care which one is selected.
3. For ordinary $n > p$ situations, if there is a high correlation between predictors, based on empirical observations, the prediction performance of LASSO is dominated by Ridge Regression.

Furthermore, Friedman et al. (2008) developed the Least Angle Regression (LARS) algorithm which is used to estimate the linear regression model [20].

Beil et al. [22], states that LARS is an efficient algorithm used because LARS has modifications to simplify LASSO calculations and produce algorithm efficiency in estimating LASSO parameters with faster computations than quadratic programming. LARS is a classical method related to the model selection method formerly known as forward selection or forward stepwise regression. Forward Selection is a method for selecting variables in linear regression, which is a technique for obtaining a regression model by selecting variables that meet specific criteria. Ransom et al. [15], the LARS algorithm is described as follows.

1. Starting with all parameter coefficients equal to zero $\beta_1, \beta_2, \ldots, \beta_k = 0$, making $\varepsilon = y$.
2. The predictor variable with the highest correlation coefficient with the remainder ε is determined.
3. The parameter coefficient β_k is estimated for x_{ik} which has the highest correlation with the remainder.
4. The remainder $\varepsilon = y - \hat{y}$ is calculated with the predictor variable x_k entered into the model.
5. The partial correlation between the remaining predictor variables and the most recent residual was calculated.
6. Steps 3 to 5 are repeated until all predictor variables are included in the model and stop when the correlation between y and x_{ik} is zero.

The LARS algorithm is a model selection method where the algorithm can be modified to be implemented in the LASSO solution. According to Zou and Hastie [23], the LARS algorithm performs an estimates of regression coefficients by taking into account both the size of the coefficients and their impact on the prediction error. The following are the steps for LASSO estimation using the LARS algorithm.

Table 4. Advantages and Disadvantages of Ridge Regression, LASSO, and Elastic Net.

Methods	Effects	Benefits
Ridge Regression	• Handles multicollinearity well by reducing coefficient variability. • Stable and consistent in providing parameter estimates. • Suitable for situations where all variables in the model are expected to contribute to the prediction.	• It does not perform variable selection automatically; all variables remain in the model with specific weights.
LASSO	• Automatic variable selection by pushing some coefficients to zero. • Produces a model that is simpler and easier to interpret.	• Sensitive to strongly correlated variables, one can choose one of several correlated variables randomly. • Model stability can be reduced if there are variables that are correlated with each other.
Elastic Net	• Combination of the advantages of Ridge (handling multicollinearity) and LASSO (variable selection). • More stable than LASSO when dealing with correlated variables. • Flexible in handling various types of data with appropriate α parameter tuning.	• Requires additional parameter tuning (α), which can affect model complexity and interpretability.

1. Finding a vector that is proportional to the correlation vector between the predictor variables and the error of each predictor variable

$$\hat{\boldsymbol{c}} = \boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_{\boldsymbol{A}})$$

Here $\hat{\boldsymbol{c}}$ is the correlation vector, $\boldsymbol{X}$ is the matrix of independent variables, $\boldsymbol{y}$ is the dependen variable and $\hat{\boldsymbol{\mu}}_{\boldsymbol{A}}$ is the mean of the predicted values.

2. Determine the largest absolute moment correlation $(\widehat{C})$,

$$\hat{C} = \underbrace{max}_{j}\{|\hat{c}_j|\}$$

so that we can obtain $s_j = sign\{\hat{c}_j\}$ for $j \epsilon A$

3. Specifying $\boldsymbol{X}_{\boldsymbol{A}}$, the set A is the active index set of the predictor variables $\{1, 2, 3, \dots, m\}$. The active index set A is determined based on the largest absolute correlation value. Defined matrix: $X_A = (\dots s_j x_j \dots)_{j \epsilon A}$ where $s_j \pm 1$, then $\boldsymbol{G}_{\boldsymbol{A}} = \boldsymbol{X}_{\boldsymbol{A}}^{\boldsymbol{T}} \boldsymbol{X}_{\boldsymbol{A}}$ and $\boldsymbol{A}_{\boldsymbol{A}} = (\boldsymbol{1}_A^T \boldsymbol{G}_A^{-1} \boldsymbol{1}_A)^{-\frac{1}{2}}$. Here $\boldsymbol{G}_{\boldsymbol{A}}$ is Gram matrix and $\boldsymbol{1}_A$ is a vector of ones corresponding to the indicates in set $\boldsymbol{A}$.

4. Calculating the value of the equiangular vector $(\boldsymbol{u}_{\boldsymbol{A}})$, the equiangular vector is a vector that divides the angles of the columns into equal measure with the angle measure less than 90°. the equiangular vector value is found using the following formula: $\boldsymbol{u}_A = \boldsymbol{X}_A \boldsymbol{\omega}_A$ with $\boldsymbol{\omega}_A = A_A \boldsymbol{G}_A^{-1} \boldsymbol{1}_A$

5. Calculating the inner product vector $(\boldsymbol{a})$:

$$\boldsymbol{a} \equiv \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{u}_{\boldsymbol{A}}$$

6. Counting $\hat{\boldsymbol{\mu}}_A$ (updating the mean)

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{A}+} = \hat{\boldsymbol{\mu}}_A + \hat{\gamma} \boldsymbol{u}_{\boldsymbol{A}}$$

where

$$\hat{\gamma} = min^{+}_{j \epsilon A^c} \left\{ \frac{\hat{C} - \hat{c}_j}{A_A - a_j}, \frac{\hat{C} + \hat{c}_j}{A_A + a_j} \right\}$$

$min^{+}_{j \epsilon A^c}$ indicates that the selected is the positive minimum value of j, which is not a set A.

7. Determine the value $\hat{\beta}$, where $\hat{\beta}$ is a candidate for the LASSO coefficient

8. Checks whether $sign(\hat{\beta}_j) = sign(\hat{c}_j) = s_j$. $\hat{\beta}_j$ and $\hat{c}_j$ have the same sign as s_j, then the selection of the following predictor variable can be continued.

9. Repeat the steps for each variable selection until all predictor variables have been selected.

In Ridge Regression, three concepts must be understood, namely regularization, l_1-norm or loss function or l_1-regularization, and l_2-loss function or l_2-regularization. Regularization is used to solve the problem of inappropriate model performance, meaning that a model has good performance for training data but has poor performance for test data. Regularization solves this problem by adding a penalty to the objective function and controlling the complexity of the model with that penalty. Regularization is usually used for situations with a large number of variables, the ratio of the number of observations and small variables, and the presence of multicollinearity. l_1-regularization minimizes the objective function by adding a penalty to the total absolute value of the coefficient commonly known as the smallest absolute deviation method, while l_2-regularization minimizes the objective function by adding a penalty to the sum of the squares of the coefficients. The points of difference between l_1-regularization and l_2-regularization can be seen in Table 1.

Meanwhile, Lasso Regression uses the - regularization technique in the objective function. The advantage of lasso regression compared to Ridge Regression is that lasso regression can choose the default variables and parameter shrinkage. Although both Ridge and LASSO Regression are used to treat multicollinearity, computationally Ridge Regression is more efficient than LASSO regression.

3.3. Elastic Net

Elastic Net (EN) first emerged as a result of criticism of LASSO, whose variable selection is so data-dependent that it is unstable [23]. EN is a selection method that combines Ridge Regression and LASSO by combining l_1-regularization and l_2-regularization on β. The combination of these two constraints is expected to balance the weaknesses of each method (Ridge and LASSO) with the l_1-regularization constraint, resulting in a simpler model due to the Shrinkage of some, which is exactly zero while at the same time l_2-regularization produces a model that does not select variables but increases the clustering and shrinking effect of β Although the two regularizations shrink the estimator β towards zero, they have different effects. Using l_2-regularization tends to produce β which is small but not zero while l_1-regularization tends to produce some regression coefficients that are exactly zero and some other regression coefficients are small. The combination of the two results in several regression coefficients of exactly zero but not as many as using only l_1-regularization.

Therefore, EN combines Ridge and LASSO regression penalties to get the best one with the aim of minimizing the loss function, which includes l_1 and l_2 penalties.

Loss function

$$L_{Enet}(\hat{\beta}) = \frac{\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2}{2n} + \lambda\left(l_2\sum_{j=1}^{m}\hat{\beta}_j^2 + l_1\sum_{k=1}^{p}|\beta_k|\right)$$

where α is the combined Ridge and Lasso regression.

3.4. The Comparison of Ridge Regression, LASSO, and Elastic Net

Ridge Regression, LASSO, and Elastic Net are popular Shrinkage modelling methods. Model shrinkage refers to statistics and machine learning techniques where regression coefficients or other model parameters are reduced or constrained from their values in the initial regression model. The primary goal of model shrinkage is to decrease variability and complexity within the model, which in turn can reduce overfitting and improve the model's generalization to unseen data. These three methods are used to increase the accuracy of predictions in the case of multicollinearity problems [24]. According to Friedman et al. [20], the comparison of the three methods in overcoming multicollinearity problems can be seen in the following Table 2.

According to Osman et al. [25], the penalty in the context of regression refers to the use of regularization techniques, such as L_1 (LASSO), L_2 (Ridge Regression), or a combination of the two (Elastic Net), which affects the regression coefficients in the model. The effect of this penalty is very important because it affects the complexity, feature selection, and generalization of the model. According to Tibshirani (1996) [8], Friedman et al. (2008) [20], and Zou & Hastie (2005) [23], the main effect of penalties on the regression coefficients of the three methods can be seen in the following Table 3.

In addition to seeing how the three methods deal with multicollinearity problems and seeing the effects and benefits of adding penalties to the regression coefficients, we can compare Ridge, LASSO, and Elastic Net regression, by evaluating the advantages and disadvantages of each method in the Table 4 [26].

Ridge Regression is effective when there are many highly correlated predictor variables, and when the primary goal is to reduce the variability of coefficients in the model without having to remove variables [20]. LASSO would be suitable for situations where feature selection is required to obtain a simpler and more interpretive model, especially when there is confidence that only a few variables influence the response [27]. Meanwhile, Elastic Net will be ideal for use when there are many correlated variables, and when it is necessary to maintain groups of correlated variables in the model while carrying out adaptive variable selection [23].

3.5. The Theoretical Applications of Ridge Regression, LASSO, and Elastic Net

Several studies comparing the three methods by applying them in several cases have been carried out. These studies provide an in-depth understanding of the conditions under which each regression method (Ridge, LASSO, or Elastic Net) is superior and how they can be used effectively in various research situations. Research conducted by Omar et al. (2023) [28] applied the Ridge Regression, LASSO, and KNN methods to evaluate the dynamic response of aluminium and ABS materials. All three methods are suitable for machine learning tasks that need to be run efficiently and do not require a lot of computing power for training and testing. This is because the algorithm is relatively simple and does not have many hyperparameters to tune, which can make it faster to train and test than some other models. These algorithms are also relatively easy to understand and interpret, making them a good choice for practitioners who are new to machine learning or do not have a strong mathematical background. Research conducted by Raouhi et al. (2022) [29] has compared Ridge, LASSO, and Elastic Net regression in the efficiency of irrigation water use under climate change conditions. The research results show that LASSO performs variable selection and parameter shrinkage, while Ridge Regression only performs parameter shrinkage and finally includes all coefficients in the model. In the presence of correlated variables, Ridge Regression may be a better choice. Moreover, Ridge Regression works best when the Least Squares estimate has a higher variance. Another research conducted by Kelachi et al. (2023) [30] used Ridge Regression, LASSO, and Ridge regression techniques to analyze heart rate data obtained from Lulu Briggs Health Center, the University of Port Harcourt, which had multicollinearity problems. The research results show that Ridge, LASSO, and Ridge Regression techniques can solve multicollinearity problems and overcome overfitting in model building. However, the choice of technique depends on the type of data being considered.

This is different from the research conducted by Kılıçoğlu and Yerlikaya-Özkurt (2024) where the Ridge Regression, LASSO, and Elastic Net methods were applied to different simulated data sets with different characteristics and also real-world data sets [24]. Based on the performance results, the methods are compared according to a multi-criteria decision-making method called TOPSIS, and a preference order is determined for each data set. The results show that LASSO and Elastic Net models outperform Ridge Regression, both on simulated and real-world datasets. Although the consistent performance of LASSO and Elastic Nets across diverse datasets suggests inherent advantages in terms of prediction accuracy, variable selection, or robustness to data properties, caution is warranted before making broad generalizations. Specific characteristics of the data set used in this research, such as sample size, predictor variables, outcome measures, and underlying data distribution, may influence the observed superiority of LASSOs and Elastic Nets.

Cleophas & Zwinderman (2015) have also compared Ridge Regression, LASSO, and Elastic Net on data from 250 patients on the patient's microarray gene expression levels and drug efficacy scores [31]. The results show that, specifically, the optimal scaling sensitivity of Ridge Regression produces more significant predictors in the data. Additionally, LASSO's optimal scaling shrinks some b values of the variable to zero, and therefore, it is suitable if you are looking for a limited number of strong predictors. On the other hand, the Elastic Net's optimal scaling performs better than the LASSO if the number of predictors is much greater than the number of observations. Another research conducted by Gilbraith et al. (2021) [32] presents four unique prediction techniques namely LASSO, Elastic Net, Partial Least Squares Regression, and Ridge Regression combined with several data pre-processing methods, utilizing various types of oil and oil peroxide value (PV) as well as incorporating natural aging for peroxide production. The results showed that although no individual regression model was be the best, the global models for each regression type and the pre-processing methods showed good agreement between all regression types when performed in their optimal scenarios. In addition, the research results show promising progress in developing a full global model for the determination of the PV of vegetable oils.

4. CONCLUSIONS

Ridge Regression, LASSO, and Elastic Net are extensions of the ordinary least squares regression aimed at addressing multicollinearity and overfitting issues. These methods apply penalties to the regression coefficients to stabilize models and improve prediction accuracy. Ridge Regression adds an L2 penalty to minimize variance and stabilize coefficient estimates, particularly effective for handling multicollinearity without eliminating any features. Ridge Regression is best suited for datasets with high multicollinearity where all features are considered necessary. LASSO introduces an L1 penalty, which can shrink some coefficients to zero, making it ideal for feature selection in cases where some predictors have little predictive power. LASSO is preferred for feature selection in sparse datasets, reducing complexity by setting some coefficients to zero. Elastic Net combines both L1 and L2 penalties, balancing feature selection and the reduction of coefficient variance. It is suitable for datasets where predictors are highly correlated. Elastic Net offers a flexible approach by combining the strengths of Ridge and LASSO, making it practical for correlated predictors that require both regularization and feature selection. Each of these methods has unique advantages, and the choice of method depends on the data structure and analysis goals.

AUTHOR INFORMATION

Corresponding Author

Dewi Retno Sari Saputro — Mathematics Study Program, Universitas Sebelas Maret, Surakarta-57126 (Indonesia);
orcid.org/0000-0002-6569-394X
Email: dewiretnoss@staff.uns.ac.id

Authors

Nugroho Lambang Wahyu — Center for Health Financing Policy and Health Insurance Management, Universitas Gadjah Mada, Yogyakarta-55281 (Indonesia);
orcid.org/0000-0002-6303-5441

Yekti Widyaningsih — Statistics Study Program, Universitas Indonesia, Depok-16424 (Indonesia);
orcid.org/0000-0002-1309-6916

Author Contributions

D. R. S. S., Y. W., and N. L. W. designed the research. D. R. S. S. performed the GWRR and Geographically Weighted-Least Absolute Shrinkage theoretically and also performed multicollinearity, H. P. performed ridge regression in shrinkage methods, D. R. S. S. and Y.W. compared - regularization and - regularization. Y. W. and N. L. W., worked on the performance for selection operator (GW-LASSO) through least square error (LSE) theoretically, D. R. S. S., Y. W., and N. L. W. wrote the paper with inputs from all authors.

Conflicts of Interest

The authors declare no conflict of interest.

ACKNOWLEDGEMENT

The authors would like to thank the Mathematical Soft Computing Research Group for their moral support and insightful comments to improve this article, especially Purnami Widyaningsih and Sutanto.

REFERENCES

[1] P. Barua, S. H. Rahman, and M. H. Molla. (2022). "Analysis of Climate Change Induced Parameters of South-Eastern Coastal Islands of Bangladesh: Comparison from 1977 to 2017". *Journal of Multidisciplinary Applied Natural Science.* **2** (1): 47-57. 10.47352/jmans.2774-3047.107.

[2] S. B. Mahfoud, H. E. Oirdi, E. H. E. Mouhab, N. Abdellahi, F. Ahmed, J. Mostafi, M. Maaroufi, S. Lotfi, K. E. Kharrim, and D. Belghyti. (2023). "Evaluation of The Integrated Protocol for The Management of Severe Malnutrition in Children at The National Hospital of Nouakchott-Mauritania". *Journal of Multidisciplinary Applied Natural Science.* **4** (1): 130-138. 10.47352/jmans.2774-3047.199.

[3] C. G. Soh and Y. Zhu. (2022). "A sparse fused group lasso regression model for fourier-transform infrared spectroscopic data with application to purity prediction in olive oil blends". *Chemometrics and Intelligent Laboratory Systems.* **224**. 10.1016/j.chemolab.2022.104530.

[4] N. T. Negero, G. F. Duressa, L. Rathour, and V. N. Mishra. (2023). "A novel fitted numerical scheme for singularly perturbed delay parabolic problems with two small parameters". *Partial Differential Equations in Applied Mathematics.* **8**. 10.1016/j.padiff.2023.100546.

[5] Q. Gao, Y. He, Z. Yuan, J. Zhao, B. Zhang, and F. Xue. (2011). "Gene- or region-based association study via kernel principal component analysis". *BMC Genetics.* **12** : 75. 10.1186/1471-2156-12-75.

[6] J. Y.-L. Chan, S. M. H. Leow, K. T. Bea, W. K. Cheng, S. W. Phoong, Z.-W. Hong, and Y.-L. Chen. (2022). "Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review". *Mathematics.* **10** (8). 10.3390/math10081283.

[7] C. W. Beaver and J. F. Harbertson. (2016). "Comparison of Multivariate Regression Methods for the Analysis of Phenolics in Wine Made from TwoVitis viniferaCultivars". *American Journal of Enology and Viticulture.* **67** (1): 56-64. 10.5344/ajev.2015.15063.

[8] R. Tibshirani. (1996). "Regression Shrinkage and Selection Via the Lasso". *Journal of the Royal Statistical Society Series B: Statistical Methodology.* **58** (1): 267-288. 10.1111/j.2517-6161.1996.tb02080.x.

[9] S. Al-Shoukry, B. J. M. Jawad, Z. Musa, and A. H. Sabry. (2022). "Development of predictive modeling and deep learning classification of taxi trip tolls". *Eastern-European Journal of Enterprise Technologies.* **3** (3 (117)): 6-12. 10.15587/1729-4061.2022.259242.

[10] Y. Xiong, W. Yang, H. Liao, Z. Gong, Z. Xu, Y. Du, and W. Li. (2022). "Soft variable selection combining partial least squares and attention mechanism for multivariable calibration". *Chemometrics and Intelligent Laboratory Systems.* **223**. 10.1016/j.chemolab.2022.104532.

[11] U. R. V. Aires, D. D. D. Silva, E. I. Fernandes Filho, L. N. Rodrigues, E. M. Uliana, R. S. S. Amorim, C. B. M. Ribeiro, and J. A. Campos. (2022). "Modeling of surface sediment concentration in the Doce River basin using satellite remote sensing". *Journal of Environmental Management.* **323** : 116207. 10.1016/j.jenvman.2022.116207.

[12] L. Wang, S. Fang, Z. Pei, D. Wu, Y. Zhu, and W. Zhuo. (2022). "Developing machine learning models with multisource inputs for improved land surface soil moisture in China". *Computers and Electronics in Agriculture.* **192**. 10.1016/j.compag.2021.106623.

[13] Ş. Çelik, T. Şengül, B. Söğüt, H. Inci, A. Y. Şengül, A. Kayaokay, and T. Ayaşan. (2018). "Analysis of Variables Affecting Carcass Weight of White Turkeys by Regression Analysis Based on Factor Analysis Scores and Ridge Regression". *Brazilian Journal of Poultry Science.* **20** (2): 273-280. 10.1590/1806-9061-2017-0574.

[14] S. C. Basak and S. Majumdar. (2015). "Prediction of Mutagenicity of Chemicals from Their Calculated Molecular Descriptors: A Case Study with Structurally Homogeneous versus Diverse Datasets". *Current Computer-Aided Drug Design.* **11** (2): 117-23. 10.2174/1871524915666150722121322.

[15] C. J. Ransom, N. R. Kitchen, J. J. Camberato, P. R. Carter, R. B. Ferguson, F. G. Fernández, D. W. Franzen, C. A. M. Laboski, D. B. Myers, E. D. Nafziger, J. E. Sawyer, and J. F. Shanahan. (2019). "Statistical and machine learning methods evaluated for incorporating soil and weather into corn nitrogen recommendations". *Computers and Electronics in Agriculture.* **164**. 10.1016/j.compag.2019.104872.

[16] C. P. Herter, E. Ebmeyer, S. Kollers, V. Korzun, T. Wurschum, and T. Miedaner. (2019). "Accuracy of within- and among-family genomic prediction for Fusarium head blight and Septoria tritici blotch in winter wheat". *Theoretical and Applied Genetics.* **132** (4): 1121-1135. 10.1007/s00122-018-3264-6.

[17] T. Kusunoki, S. Hatanaka, M. Hariu, Y. Kusano, D. Yoshida, H. Katoh, M. Shimbo, and T. Takahashi. (2022). "Evaluation of prediction and classification performances in different machine learning models for patient -specific quality assurance of head-and-neck VMAT plans". *Medical Physics.* **49** (1): 727-741. 10.1002/mp.15393.

[18] A. Anjum, A. A. Shaikh, and N. Tiwari. (2023). "Experimental investigations and modeling for multi-pass laser micro-milling by soft computing-physics informed machine learning on PMMA sheet using CO2 laser". *Optics & Laser Technology.* **158**. 10.1016/j.optlastec.2022.108922.

[19] R. Carvalheiro, E. C. Pimentel, V. Cardoso, S. A. Queiroz, and L. A. Fries. (2006). "Genetic effects on preweaning weight gain of Nelore-Hereford calves according to different models and estimation methods". *Journal of Animal Science.* **84** (11): 2925-33. 10.2527/jas.2006-214.

[20] J. Friedman, T. Hastie, and R. Tibshirani. (2008). "Sparse inverse covariance estimation with the graphical lasso". *Biostatistics.* **9** (3): 432-41. 10.1093/biostatistics/kxm045.

[21] A. Zöngür and M. A. Buzpinar. (2023). "AI-assisted antifungal discovery of Aspergillus parasiticus and Aspergillus flavus: investigating the potential of Asphodelus aestivus, Beta vulgaris, and Morus alba plant leaf extracts". *Modeling Earth Systems and Environment.* **9** (2): 2745-2756. 10.1007/s40808-022-01658-2.

[22] C. T. Beil, V. A. Anderson, A. Morgounov, and S. D. Haley. (2019). "Genomic selection for winter survival ability among a diverse collection of facultative and winter wheat genotypes". *Molecular Breeding.* **39** (2). 10.1007/s11032-018-0925-8.

[23] H. Zou and T. Hastie. (2005). "Regularization and Variable Selection Via the Elastic Net". *Journal of the Royal Statistical Society Series B: Statistical Methodology.* **67** (2): 301-320. 10.1111/j.1467-9868.2005.00503.x.

[24] Ş. Kılıçoğlu and F. Yerlikaya-Özkurt. (2024). "A novel comparison of shrinkage methods based on multi criteria decision making in case of multicollinearity". *Journal of Industrial and Management Optimization.* **20** (12): 3816-3842. 10.3934/jimo.2024072.

[25] H. Osman, M. Ghafari, and O. Nierstrasz. (2017). "Automatic feature selection by regularization to improve bug prediction accuracy". presented at the 2017 IEEE Workshop on Machine Learning Techniques for Software Quality Evaluation (MaLTeSQuE). 10.1109/MALTESQUE.2017.7882013.

[26] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor. (2023). "An Introduction to Statistical Learning". 10.1007/978-3-031-38747-0.

[27] C. J. Greenwood, G. J. Youssef, P. Letcher, J. A. Macdonald, L. J. Hagg, A. Sanson, J. McIntosh, D. M. Hutchinson, J. W. Toumbourou, M. Fuller-Tyszkiewicz, and C. A. Olsson. (2020). "A comparison of penalised regression methods for informing the selection of predictive markers". *PLoS One.* **15** (11): e0242730. 10.1371/journal.pone.0242730.

[28] I. Omar, M. Khan, and A. Starr. (2023). "Suitability Analysis of Machine Learning Algorithms for Crack Growth Prediction Based on Dynamic Response Data". *Sensors (Basel).* **23** (3). 10.3390/s23031074.

[29] E. M. Raouhi, M. Lachgar, and A. Kartit. (2022). In: "World Integrated Trade Solution 2020, (Lecture Notes in Electrical Engineering, ch. Chapter 22". 233-240. 10.1007/978-981-33-6893-4_22.

[30] K. Enwere, E. Nduka, and U. Ogoke. (2023). "Comparative Analysis of Ridge, Bridge and Lasso Regression Models In the Presence of Multicollinearity". *IPS Intelligentsia Multidisciplinary Journal.* **3** (1): 1-8. 10.54117/iimj.v3i1.5.

[31] A. Rajkomar, J. Dean, and I. Kohane. (2019). "Machine Learning in Medicine". *The New England Journal of Medicine.* **380** (14): 1347 -1358. 10.1056/NEJMra1814259.

[32] W. E. Gilbraith, J. C. Carter, K. L. Adams, K. S. Booksh, and J. M. Ottaway. (2021). "Improving Prediction of Peroxide Value of Edible Oils Using Regularized Regression Models". *Molecules.* **26** (23). 10.3390/molecules26237281.