ISSN 2303-4521
Periodicals of Engineering and Natural Sciences *Original Research*
Vol. 11, No. 2, March 2023, pp.36-47

A comparative study between shrinkage methods (ridge-lasso) using simulation

Zainab Fadhil Ghareeb[1], Suhad Ali Shaheed AL-Temimi[2]
[1] Department of Statistics, College of Administration and Economics, Mustansiriyah University, Baghdad, Iraq.
[2] Department of Statistics, College of Administration and Economics, Mustansiriyah University, Baghdad, Iraq.

ABSTRACT

The general linear model is widely used in many scientific fields, especially biological ones. The Ordinary Least Squares (OLS) estimators for the coefficients of the general linear model are characterized by good specifications symbolized by the acronym BLUE (Best Linear Unbiased Estimator), provided that the basic assumptions for building the model under study are met. The failure to achieve one of the basic assumptions or hypotheses required to build the model can lead to the emergence of estimators with low bias and high variance, which results in poor performance in both prediction and explanation of the model in question. The hypothesis that there are no multiple linear relationships between the explanatory variables is considered one of the leading hypotheses on which the model is based. Thus, the emergence of this problem leads to misleading results and high (Wide) confidence limits for the estimators associated with those variables due to problems characterizing the model. Shrinkage methods are considered one of the most effective and preferable ways to eliminate the multicollinearity problem. These methods are based on addressing the multicollinearity problems by reducing the variance of estimators in the model. Ridge and Lasso methods represent the most and most common of these methods of shrinkage. The simulation was carried out for different sample sizes (40, 120, 200) and some variables (P=30, 60) in the first and second experiments arbitrarily and at the level of low, medium, and high correlation coefficients (0.2, 0.5, 0.8). When (p=30, 60) Lasso method has the smallest (MSE) than the Ridge method. The Lasso method proved its efficiency by obtaining the least MSE. Optimal Penalty parameter (λ) chosen from Cross-Validation through minimizing (MSE) of prediction. We see a rapid increase for (MSE) for both (Ridge-Lasso) where the top axis indicates the number of model variables, and when the correlation between variables increases and sample size too, we can see the (MSE) values increase in the Ridge method than the Lasso method. A ridge method gives greater efficiency when the sample size is more significant than variables (p<n), but the Ridge method cannot shrink coefficients to precisely zero. So, the elasticity of ridge coefficients decreases, but variance increases bias, also (MSE) first remains relatively constant and then increases fast.

Keywords: The general linear model, multicollinearity, Shrinkage Methods, Ridge, Lasso, and penalty parameter.

Corresponding Author:

Zainab Fadhil Ghareeb
Department of Statistics,
College of Administration and Economics,
Mustansiriyah University, Baghdad, Iraq.
E-mail: badeaa99@uomustansiriyah.edu.iq

1. Introduction

The development of regression parameter estimation began in 1795 when Guss proposed the Ordinary Least Squares method, later published in 1805 by Adrine-Marie - Legendre. In 1922, Fisher Presented the Maximum likelihood method, characterized by a set of characteristics such as consistency, sufficiency, and efficiency (BLUE) [1-3].

Using the Ordinary Least Squares (OLS) method to estimate the parameters of a general linear regression model to obtain a suitable model for the prediction can sometimes lead to wrong decisions and results (misleading) because of the lack of basic assumptions required to build and analyze regression models. This has necessitated, therefore, that researchers seek to find other methods of higher accuracy, which are often modifications made in one way or another to the method (OLS) [4].

The real relationship between the response variable (Y) and the explanatory variables (X_1, X_2,…X_p) if it is almost linear, then the OLS estimates will have a low bias, and if n>p (the size of the observations is much greater than the number of variables in the model), it also tends to have a low variation. But if n is not much more significant than P, there may be a considerable variation in the estimates, which leads to overfitting and weak predictive power of the model. On the other hand, if p>n, the OLS solution is not a single solution, but there is more than one solution to estimate the parameters [5-7].

In short, shrinkage is a regulation method that involves fitting a regression model using all p predictors under some constraint on the size of their estimated coefficients.

The essential characteristics of this method can be summarized in the following points:

- Reducing the variability of the estimates means improving the model's stability.
- Setting some of the coefficients to zero allows for variable selection.

The shrinkage method is considered one of the most effective and preferable ways to reduce the multicollinearity problem. It addresses that problem by reducing the variance of estimators in the model.

This paper will highlight three estimation methods: Ridge and Lasso [2, 8].

In practical application, researchers face several problems, especially when the number of explanatory variables affecting the response variable (dependent) is large or close to the number of observations, as well as when these variables are correlated to each other (multicollinearity problem) and the consequent adverse effects of these problems on future predictions, Accordingly, the problem of this study is to put forward methods that contribute to shrinking these variables and thus shrinking the parameters of the regression model to reach an optimal model that has excellent and high explanatory and predictive power—reaching a statistical model that better represents the data through shrinkage methods. The purpose of using the shrinkage methods (Ridge and Lasso) is embodied in accessing the selection of the significant explanatory variables in the model under study. To determine the best shrinking method at different sample sizes and correlation coefficient values, a comparison is made between the results of the three methods using the mean square error criterion when there is a multicollinearity problem.

2. Literature review

Several studies and research discussed the linear relationship between the explanatory variables, methods of detection, and the extent of their impact on the results of the analysis, as well as ways to treat those using modern statistical methods.

In 1962, Horel presented several concepts that formed the basis of the Ridge Regression (RR) methodology. Then in 1970, Hoerl and Kennard presented an article considered a significant development of the RR method. The most important part of that article was focused on the ordinary RR estimator $b_{RR} = (X`X + kI)^{-1}X`y$, where k is an exogenous parameter that has to be determined. Also, the article showed that there is always a k>0 such that the mean squared error of the (b_{RR}) method is less than the mean squared error of the (b_{OLS}) method [1].

In 2018, the two researchers, Al-Hassan, Y.M., and Al-Kassab, studied the use of the RR method to determine the best value of the constant k and choose it to eliminate the problem of multicollinearity between the explanatory variables. In this context, the two researchers worked on increasing the sample size, dropping or deleting some variables with high correlation and finding the parameters of the linear regression model to represent the best model using biased estimation methods (PCR, RR). The results showed that the (RR) method gave the best parameter K and the highest efficiency as well for the estimators [3]. In 2017, the researcher (Fatima Assim Mahdi) compared the RR and PC methods using simulation and application on real data in her thesis. The thesis mentioned above proposed a method for testing the bias parameter ($\hat{k}$) by modifying the ($\hat{k}_{HKB}$) method, and the proposed value gave an excellent performance in terms of reducing MSE when there is high multicollinearity between the explanatory variables and using the standard ridge deviation estimators [8].

M. Goldsmith and A.F.M Smith proposed a new derivation for the estimator RR and its generalization. In 1977, Richard F. Gunst and Robert L. Mason used MSE (Mean Squared Error) criterion to compare five estimators of regression coefficients (least squares, principal components, ridge regression, latent root, and shrunken estimator). In this context, each of the biased estimators displayed improvement in mean squared error over least squares for a wide range of options of the parameters of the model [11-14]. Also, the results of a simulation encompassing all five estimators pointed out that the principal components and latent root estimators perform best overall, while the ridge regression estimator has the possibility of a minor mean square error than either of these [11, 15].

In 2013, a study by Irfan and Javed compared Partial Least Squares Regression with other prediction methods (Ordinary Least Squares, Ridge Regression and Principal Components Regression) to address the problem of multicollinearity in (GDP) data of Pakistan. Also, the study compared all prediction methods for efficiency by using RMSE and found that Partial Least Squares Regression (PLSR) provides better prediction as compared to the other prediction methods [15].

Robert TIBSHIRAN proposed a new method of estimation in linear models called the Lasso short (Least absolute shrinkage and selection operator). This method minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. As a result of the nature of this constraint, it tends to produce some coefficients that are precisely (0) and hence gives interpretable models. Simulation studies conducted by TIBSHIRANI indicated that the Lasso method is characterized by some favourable properties of subset selection and Ridge Regression. Also, the Lasso method is considered quite general and can be used or applied in various statistical models [16-20].

3. Method

3.1. General linear model (GLM)

The general linear model is a procedure by which a single response variable(y) is represented by a combination of explanatory variables X_1 ،X_2 ،..., Xp, and the model can be described as follows [7, 10]:

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon} \qquad \dots (1)$$

$\underline{y}$: ($n \times 1$) vector representing the response variable (dependent).

X: A total-ranked matrix with dimensions (n×p) defined for explanatory variables.

$\underline{\beta}$:A vector whose dimension is (p×1) for the undefined regression parameters or coefficients.

$\underline{\varepsilon}$:A vector with dimensions (n×1) of random errors achieving, E (ε) = 0 and V(ε) = $\sigma^2{}_{In}$.

I_n: Identity matrix with (n$\times$n).

The coefficient of the general linear model in equation (1) is estimated using the Ordinary Least Squares (OLS) method, the most common estimation method based on minimizing the sum of squared deviations, as shown below.

$$\varepsilon`\varepsilon = Y`Y - 2\,\beta`\,X`Y + \beta`\,X`\,X\beta \qquad \dots (2)$$

And when we take the derivative of equation (ε` ε) concerning parameter (β) according to the method of least squares, we get the OLS estimator:

$$b_{ols} = (X`\,X)^{-1}\,X`\gamma \qquad \dots (3)$$

3.2. The collinear relationship between explanatory variables

One of the most important assumptions of the General Linear Model (GLM) analysis is that the explanatory variables are not highly correlated. The interpretation of the regression coefficient is usually the amount of change in the response variable when an explanatory variable is increased by one unit with all other explanatory variables held constant. This interpretation becomes invalid when there is high collinearity between the explanatory variables. On the other hand, when this collinearity is not present, the explanatory variables are considered orthogonal. However, the lack of orthogonality is not so important as to cause the analysis to be inaccurate. The absence of orthogonality with a high degree is also attributed to the problem of the data's linearity or multicollinearity. Accordingly, it is necessary to notice the presence or absence of the multicollinearity problem between the explanatory variables when building the model. [6][12]

3.3. Shrinkage method

The estimator of the Ordinary Least Squares method dominated studies and research for a long time until it was proven to be ineffective in the case of the existence of a multicollinearity problem between the explanatory variables. To understand the methodology of the shrinkage estimators, we assume that we have a matrix of explanatory variables **X** with a dimension (n×p) and that the vector of the response variable is **y** with a dimension (n×1). In addition, we assume that the sample means have been excluded from the data set under study and, therefore, **1**'x = 0 and **1**'y=0, where **1** is a vector (n×1) of ones. Also, the **X** matrix will be decomposed to get a more comprehensive understanding of the shrinkage estimators, and thus the X matrix can be written as follows: [8][15]

$$X = H\Lambda^{\frac{1}{2}}\, G` \qquad \dots (4)$$

H: represents an (n×p) matrix that fulfils the orthogonality condition H`H=**I_p**.

$\Lambda^{\frac{1}{2}}$: represents a (p×p) diagonal matrix of ordered singular values of X, that is $\lambda_1^{\frac{1}{2}} \geq \lambda_2^{\frac{1}{2}} \geq \cdots . \lambda_p^{\frac{1}{2}} \geq 0$ which represents the Eigenvalues and so that (β) is estimable.

G: represents a (p×p) orthogonal matrix, and its columns are the Eigen Vectors of the X`X matrix.

The Information Matrix X`X can be written in the following form:

$$X`X = G\ \Lambda^{\frac{1}{2}}\, H`H\ \Lambda^{\frac{1}{2}}\, G` = G\ \Lambda\ G` \qquad \dots (5)$$

Therefore, the least squares estimator can be rewritten as:

$$b_{ols} = (X`X)^{-1}\, XY = (G\Lambda G`)^{-1} G\ \Lambda^{\frac{1}{2}}\, H`Y$$

$$b_{ols} = G\ \Lambda^{-1}\ G`G\ \Lambda^{\frac{1}{2}}\, H`Y = G\Lambda^{\ -\frac{1}{2}}\ H`Y = GC \qquad \dots (6)$$

Where the Vector $C = \Lambda^{\frac{-1}{2}}\ H`y$ represents the vector of uncorrelated components of b_{ols}. This leads to the following:

$$E(C) = E(G`\ b_{ols}) = G`\beta = \gamma(say) and\ var\ (C) = var\ (G`b_{ols})$$

$$\therefore E(C) = G`var(b_{ols})G \qquad \dots (7)$$

So,

$$Var(C) = \sigma^2 G`(G\ \Lambda\ G`)^{-1}G = \sigma^2 G`G\ \Lambda^{-1}G`G = \sigma^2\Lambda^{-1} \quad . \qquad \dots (8)$$

It represents a diagonal matrix, and therefore the components of C are uncorrelated due to the off-diagonal elements of $Var(C)$, which represent the covariance values equal to zero. The shrinkage estimators are denoted by $b_{SH}$and their general form is:[8,16]

$$b_{SH} = G\Delta C = \sum_{i=1}^{b} \overline{g_j}\, \delta_j C_j \qquad \dots (9)$$

Where:

$\overrightarrow{g_j}$: represents the jth column of the matrix G.

δ_j: represents the jth diagonal elements of the shrinkage factors matrix Δ, and the range of the shrinkage factors is usually restricted and be: $0 \leq \delta_j \leq 1. j = 1.2. \dots . C_j$.In this research, several methods will be presented to address the problem of multicollinearity of the GLM model, the most important of which are:

3.4. Ridge method (RM)

In 1970, researchers Hoerl and Kennard introduced the Ridge Estimator (RE) method, which became one of the most popular methods for solving the multicollinearity problem of the linear regression model. This method is implemented by adding a small positive amount with a specified value between zero and one ($0 < \lambda < 1$) to the diagonal elements of the data matrix (X`X) to obtain more accurate estimators. Also, the (RR) method leads to decorrelation between explanatory variables. The normal equation below is used to calculate the estimators for the values of parameter β after the response variable y, and the explanatory variables are converted to the standard form [16, 21]:

$$\left(X`X + \lambda I_p\right)\beta = X`y \qquad \dots (10)$$

Where λ represents a nonnegative constant value (representing the bias parameter or regularization parameter) and is chosen by the analyst according to several criteria developed by Hoerl and Kennard, and I_p represents the unity matrix with dimensions (p×p).

The solution to equation (10) gives the Ridge Estimator (b_{RE}) as follows:

$$b_{RE} = \left(X`X + \lambda I_p\right)^{-1} X`y \qquad \dots (11)$$

Often, we can rewrite ridge regression in the Lagrangian form: [19]

$$b_{RE} = argmin_{\beta_0,\beta} \left\{\frac{1}{2n} \, ||y - X\beta||_2^2 + \lambda||\beta||_2^2\right\} \qquad \dots (12)$$

It is easy to prove that the ridge regression estimator in equation (11) is one of the classes of shrinkage estimators, as follows:

$$b_{RE} = \left(X`X + \lambda I_p\right)^{-1} X`y = [G(\Lambda + \lambda I)G`]^{-1} G\Lambda^{\frac{1}{2}} H`y$$

$$b_{RE} = G\left(\Lambda + \lambda I_p\right)^{-1} G`G\Lambda^{\frac{1}{2}} H`y$$

$$b_{RE} = G\left(\Lambda + \lambda I_p\right)^{-1} \Lambda^{\frac{1}{2}} H`y$$

$$b_{RE} = G\left[\left(\Lambda + \lambda I_p\right)^{-1} \Lambda\right] \Lambda^{-\frac{1}{2}} H`y$$

$$b_{RE} = G\Delta C \qquad \dots (13)$$

Where Δ =(Λ +$\lambda I)^{-1}$ Λ, which is a diagonal matrix, and j-th is the diagonal element of the matrix, Δ has the form:

$$\delta_j = \frac{\lambda_j}{\lambda_j + \lambda} \ . \quad j = 1,2,\dots,p$$

Where λ_j represents the j-th element (Eigen value) of the diagonal matrix Λ with dimension (p× p), and λ represents the ridge (regularization) parameter [13].

It can be seen, from equation (11), that the ridge estimator is a biased method, and therefore the ridge estimator for the coefficients ($\hat{b}_{RE}$) is biased, and the variance and covariance matrix for the ($\hat{b}_{RE}$) estimator is as follows [8, 18]:

$$Var(b_{RE}) = \sigma^2 (X'X + \lambda I_p)^{-1} X'X\left(X'X + \lambda I_p\right)^{-1}$$

The residual sum of the square can be formulated as follows:

$$SSE(\lambda) = (y - Xb_{ols})'(y - Xb_{ols}) + (b_{RR} - b_{ols})'X'X(b_{RR} - b_{ols}) \qquad \dots (14)$$

The total mean squared error (Total MSE) can be formulated as follows:

$$Total\ MSE(b_{RR}) = E(b_{RR} - \beta)'(b_{RR} - \beta)$$

Using the linear algebra theory of the quadratic formula as follow:

$$E(y'Ay) = tr(A)\Sigma + \mu' A\mu \qquad \dots (15)$$

$$Total\ MSE(b_{RR}) \ = \sigma^2 tr \left[\left(X'X + \lambda I_p\right)^{-1} X'X\left(X'X + \lambda I_p\right)^{-1}\right] + \beta'\left(W - I_p\right)'\left(W - I_p\right)\beta \quad \dots (15)$$

Where $W = \left[I_p + \lambda(X'X)^{-1}\right]^{-1}$, By simplifying the equation (15), we obtain:

$$Total\ MSE(b_{RR}) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + \lambda)^2} + \lambda^2 \beta'(X'X + \lambda I_p)^{-2}\beta \qquad \dots (16)$$

One of the essential points that must be considered in the Ridge method is the value of the bias parameter (λ). Several methods are proposed to find the best value of (λ) [21]. When the value of (λ =0), the ridge estimators are equal to the least squares estimators, and when (λ >0), the ridge regression estimators tend to stabilize at a specific value relative to changes in the data, but there is a bias.

In addition, the mean squares error (MSE) of the ridge method is lower than the MSE of the least square method, so the amount of bias is accepted against the reduction of the variance of the estimators [20]:

$$Var(T) \geq \frac{[1+B(\theta)]^2}{E\left[\frac{\partial lnL}{\partial\theta}\right]^2} \qquad \ldots (17)$$

Where B (θ)is the bias of the estimator.

On this basis, $b_{RE}$is the best linear estimator for β. Also, the ridge regression estimator is the Bayes estimator when the coefficients for the prior distribution are Gaussian distribution and represent the least squared error of the Penalty Function as a result of minimizing the objective function according to the Formula:

$$(y - X\beta)`(y - X\beta) + \lambda(\beta`\beta - c) \qquad \ldots (18)$$

There are several ways to test the bias parameter λ. There are several ways to test the bias parameter λ. The most popular is the Hoerl-kennard-Baldwin method, according to the following formula:

$$\hat{\lambda}_{HKB} = \frac{P\hat{\sigma}^2}{\sum_{i=1}^{p}(\hat{\beta}_{OLS}^2)\,2} \qquad \ldots (19)$$

And that $\hat{\beta}$ and $\hat{\sigma}^2$ are obtained by the method of OLS

$$\hat{\sigma}^2 = \frac{y`y - \beta'_{OLS}X'y}{n-p}$$

3.5. Least Absolute shrinkage and selection Operator (LASSO)

Robert Tibshirani proposed the Lasso method in 1996. The essence of this method is based on minimizing the Sum of Squared Errors (SSE) with attention to the sum of the absolute value of the coefficients being less than a constant. Due to the nature of this constraint, it tends to generate some coefficients equal to (0). In return, the variables greater than zero are determined after reduction and adopted as part of the model, contributing to minimizing the prediction error. This method is of great importance in addressing the problem of multicollinearity between explanatory variables [20].

The widespread use of the lasso method in the treatment of estimation problems is due to its statistical accuracy in the prediction and selection of explanatory variables and its computational accuracy. But on the other hand, the Lasso estimator is unstable when the number of explanatory variables is higher than the number of observations. Also, if there is high multicollinearity among the explanatory variables, ridge regression dominates the Lasso in prediction performance.

To address this problem in the general linear model in equation (1), the variables X_1, X_2,…, X_p are converted to the standard Formula (standardization), and therefore E(X)=0 and Var(X)=1, and let $\hat{\beta} = \left(\hat{\beta}_1 . \hat{\beta}_2 \ldots . \hat{\beta}_p\right)^T$, So the lasso estimator represents the solution as in the following equation [2, 5, 17]:

$$b_{Lasso} = \underset{\beta}{\text{argmin}}(y - X\beta)^{`}(y - X\beta) \;\; ; \qquad s.t \sum_{j=1}^{p-1}\left|\beta_j\right| \leq t \qquad \ldots (20)$$

Where t represents a tuning parameter and controls the amount of shrinkage applied to the estimates when $t \geq 0$, to determine t, let b_{ols} be the full least squares estimates and $t = \sum\left|\hat{b}_{OLS}\right|$, and then, values of $t < \lambda_{OLS}$ will cause shrinkage of the solutions towards (0), Then equation (20) can be reformulated according to Lagrange's Formula to become as follows:

$$b_{LASSO} = \min_{\beta}\left\{(y - X\beta)'(y - X\beta) + \lambda\sum_{j=1}^{q}\left|\beta_j\right|\right\} \quad ; \; q = 1,2,\ldots,p-1 \quad \ldots (21)$$

Where λ ≥0 represents the penalty parameter controlling the shrinkage amount and $\lambda\sum_{j=1}^{p}\left|\beta_j\right|$represents the penalty term (norm L_1). The penalty parameter (λ) controls the strength of the penalty function for the coefficients of the general linear model. For every $t > 0$, there is $\lambda > 0$, and therefore equations (20) and (21) become identical for any given value when $\lambda \in [0.\infty)$, then both formulas have a solution and vice versa.

In equation (21), penalty parameter λ (which is always positive(nonnegative)) is multiplied by the penalty function L$_1$ of vector $(\hat{\beta}_1.\hat{\beta}_2 \ldots.\hat{\beta}_{p-1})$. Therefore some coefficients are estimated to be zero when the value of λ is significant, while minimal values of λ can lead to mischaracterization—as in the case of ridge regression, choosing different values for λ results in different estimators of the parameters vector $\hat{\beta}_{Lasso} = \left(\hat{\beta}_1.\hat{\beta}_2 \ldots.\hat{\beta}_{p-1}\right)^T$,which is why it is so important to choose an appropriate value for the regularization parameter.

It is important to indicate that choosing the value of the regularization parameter λ for the Lasso estimator is critical because minimal values of λ can lead to a misdescription of the model, meaning that the model can tend to describe errors in the data. The results become close to the results of estimators of the OLS method. In both cases mentioned for the value of λ a high variance value will be obtained when applied to the data under study. The parameter λ in equation (21) is estimated through the Cross-Validation technique to find a suitable Value for that parameter.[9][14]

The suitable or appropriate value of parameter λ means that value that contributes to predicting the values of the response variable with the highest possible accuracy (less variance). To perform the Cross-validation technique, the data set is first divided into two subsets or folds (Say Q) of approximately equal size. The first set is called the training set, and the second is called the test set. Next, the training set is used to compute coefficient estimates, and these estimates are then validated by the test set where the fold is assigned to one of the folds while the remaining Q – 1 folds together form the training set. Usually, the number of folds Q is set to be equal to 5 or 10.

Then a grid of $\lambda = [\lambda_s]$ values is chosen, and model coefficients are computed for all λ_s values according to which the Residual Sum Squares (RSS) are calculated:

$$RSS_{\lambda s}^{q} = \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p-1} b_j(q.\lambda_s)x_{ij}\right)^2 \quad \ldots (22)$$

Where q represents the list of folds selected as a test set, the average RSS values for all folds can be obtained according to the following:

$$C.V(\lambda) = MSE_{\lambda s} = \frac{1}{Q}\sum_{q=1}^{Q} RSS_{\lambda s}^{q} \quad \ldots (23)$$

Then a value of λ equal to which provides the lowest MSE

$$\hat{\lambda}_{min} = \min CV(\lambda) \quad \ldots (24)$$

3.5. Simulation procedure

The simulation method has been used to compare shrinkage methods (Ridge-Lasso) clarified in this study to reach the best method among the applied estimation methods. The comparison, as mentioned earlier, was made by relying on the Least Mean Squares Error (MSE) criterion. In the same context, the stages of building a simulation can be summarized as follows:

1- The sample size (n) includes three sample sizes (n=40, n=120, n=200).

2- Determine the value of the correlation coefficient, which includes three levels low, medium and high so (ρ = 0.2, 0.5, 0.8).

3- Determine the mean and variance value of the generated variables ($\mu = 5, \sigma^2 = 2$).

5- Determine the repetition value for each experiment (500).

6- Determine the initial values for the parameters: For the first experiment , $\beta = (3.2,1.8,0,0,2.2,2,0,\ldots,0)$; For the second experiment, $\beta = (2.8,2.9,2.9,3,3,3,3,0 \ldots,0)$.

7- The explanatory variables are generated assuming they have a normal distribution $x_i \sim N(\mu, \sigma^2)$. As for the error, it is typically distributed with a mean of zero and a Variance of (1), $e_i \sim N(0,1)$.

8- The dependent variable was generated through the following equation:

$$y_i = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 + b_7x_7 + b_8x_8 + b_9x_9 + b_{10}x_{10} + \cdots$$

As for the comparison criteria, and as we explained above, the (MSE) criterion was relied upon, which represents the predicted error for the estimators, according to the following Formula:

$$MSE\left(\hat{\beta}^*(\alpha)\right) = \frac{1}{n}\left(y_{new} - X_{new}\hat{\beta}^*(\alpha)\right)'\left(y_{new} - X_{new}\hat{\beta}^*(\alpha)\right) \quad \ldots (36)$$

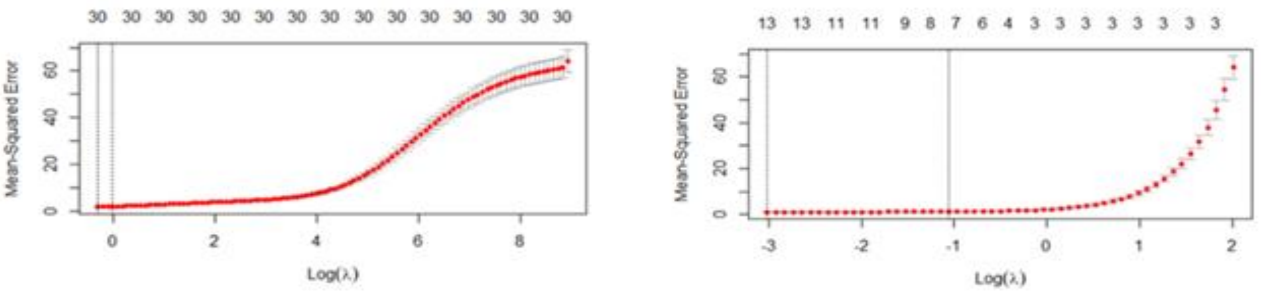
X_{new} and Y_{new} are good observations that were not used to estimate parameters $\hat{\beta}^*(\alpha)$. Simulations and model estimation were performed using R software.

4. Results and discussions

Data were generated with three levels of correlation coefficient value (low = 0.2, medium = 0.5 and high= 0.8). Two shrinkage methods (Ridge Method (RM), and Least Absolute shrinkage and Selection Operator (Lasso)) were used to address the multicollinearity problem with small, medium, and large sample sizes (40,120,200) and at two sets of several explanatory variables (30, 60) with (500) repetition. The best shrinkage parameter (λ) was selected through the least standard MSE at different sample sizes and correlation levels for each shrinkage method. As a result, the following tables show the estimators of the parameters for the explanatory variables in the general linear model. Table 1, it is clear to us the results of the simulation experiment when several variables (p = 30) with sample size (n = 40, 120, 200) and correlation coefficient (0.2, 0.5, 0.8) for (500) repetition. The Lasso method has the smallest (MSE) than the Ridge method. The Lasso method proved its efficiency by obtaining the least MSE. Optimal Penalty parameter (λ) chosen from Cross-Validation through minimizing (MSE) of prediction. Mean squared error (MSE) remains relatively constant and increases fast. In the Ridge Method, the penalty parameter (λ) increases, and the elasticity of ridge coefficients decreases, but variance increases bias. Also (MSE) first remains relatively constant and then increases fast.

Table 1. (RE, Lasso) with (MSE) when (p=30) and best Penalty parameter λ

Sample Size	**Correlation Coefficient**	**MSE** **Best Penalty parameter**	
		RM	**Lasso**
40	**0.2**	0.256398 $\lambda = \mathbf{0.475144}$	0.040111 $\lambda = \mathbf{0.00438}$
	0.5	0.354806 $\lambda = \mathbf{0.758623}$	0.0504465 $\lambda = \mathbf{0.00441}$
	0.8	0.439002 $\lambda = \mathbf{0.821458}$	0.051536 $\lambda = \mathbf{0.002523}$
120	**0.2**	0.644263 $\lambda = \mathbf{0.471331}$	0.5478987 $\lambda = \mathbf{0.053954}$
	0.5	0.750828 $\lambda = \mathbf{0.66546}$	0.5546814 $\lambda = \mathbf{0.07476}$
	0.8	0.922708 $\lambda =$ $\mathbf{0.806107}$	0.511579 $\lambda = \mathbf{0.025197}$
200	**0.2**	0.722982 $\lambda = \mathbf{0.444735}$	0.619208 $\lambda = \mathbf{0.042452}$
	0.5	0.83579 $\lambda = \mathbf{0.581336}$	0.615111 $\lambda = \mathbf{0.042418}$
	0.8	1.053697 $\lambda = \mathbf{0.80928}$	0.600951 $\lambda = \mathbf{0.017339}$



Ridge Method

Lasso Method

Figure 1. The fast increase for (MSE) for both (Ridge-Lasso), where the Top axis indicates the number of model variables

Figure 1 shows the fast increase for (MSE) for both (Ridge-Lasso), where principal axis indicates the number of model variables—the number of variables p=30, sample size n=200 with high correlation.

When the correlation between variables increases and the sample size too, we can see the (MSE) values increase in the Ridge method than in the Lasso method, as in Figure 2.

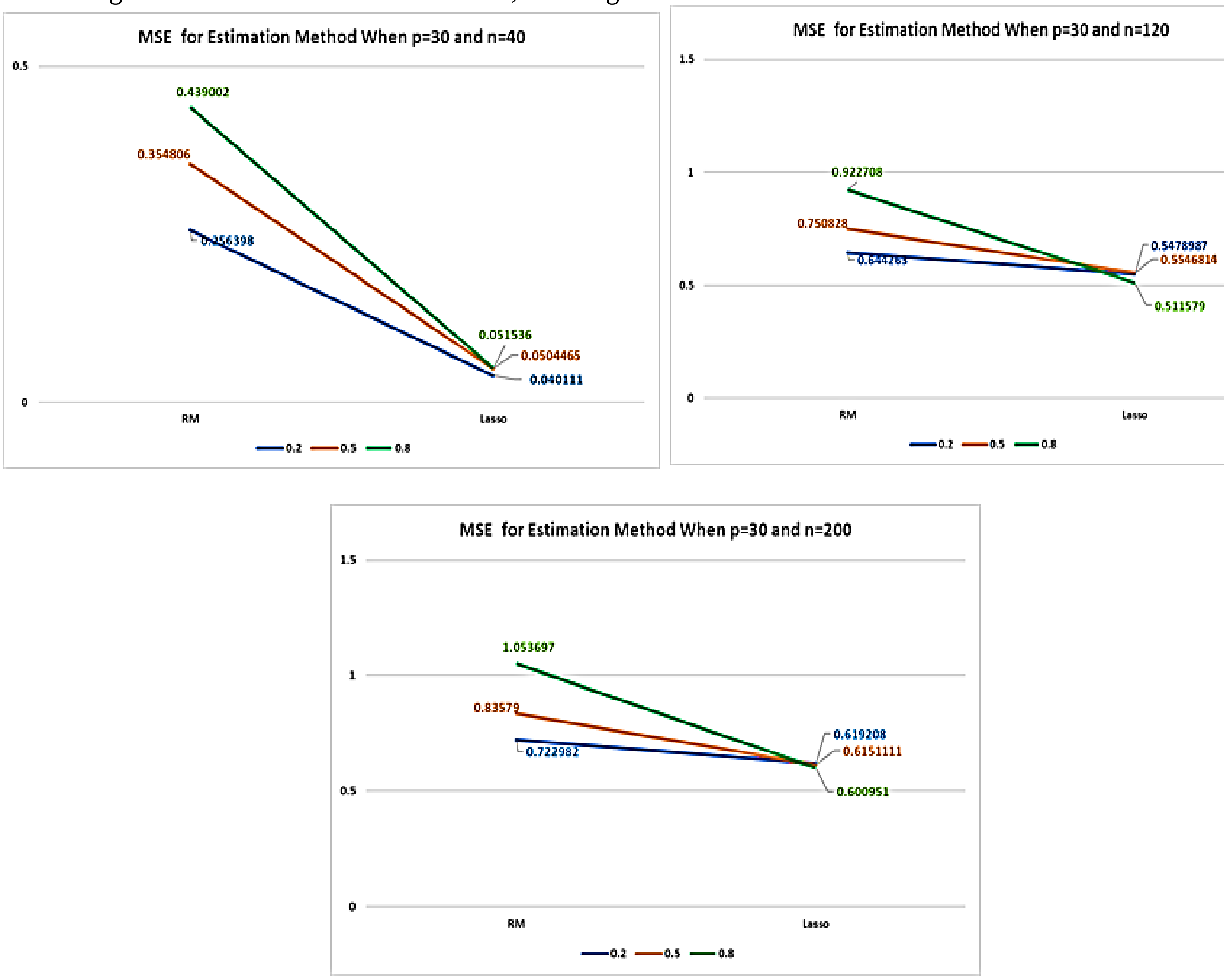


Figure 2. Comparisons between (Lasso-Ridge) via MSE

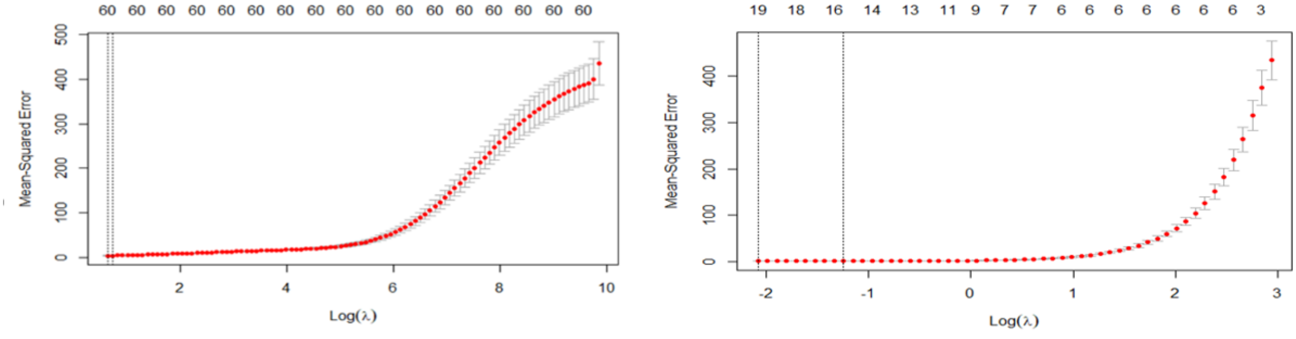
Table 2 shows the results of the simulation experiment when the number of variables (p = 60) with sample size (n = 40, 120, 200) and correlation coefficient (0.2, 0.5, 0.8) for (500) repetition. The Lasso method still has most minor (MSE) than the Ridge method. The optimal Penalty parameter (λ) is also chosen from Cross-Validation through minimizing (MSE) of prediction. Mean squared error (MSE) remains relatively constant and increases fast. In the Ridge Method penalty parameter (λ) increases when the sample size is less than several variables, a ridge method gives greater efficiency when the sample size is more significant than variables (p<n). The elasticity of ridge coefficients decreases, but variance increases bias. Also (MSE) first remains relatively constant and then increases.

Table 2. (RE, Lasso) with (MSE) when (p=60) and best Penalty parameter λ

Sample Size	**Correlation Coefficient**	**MSE Best Penalty parameter**	
		RR	**Lasso**
40	**0.2**	42.08362 $\boldsymbol{\lambda = 100.1224}$	0.142217 $\boldsymbol{\lambda = 0.316008}$
	0.5	33.96404 $\boldsymbol{\lambda = 138.2421}$	0.316008 $\boldsymbol{\lambda = 0.18522}$
	0.8	19.15464 $\boldsymbol{\lambda = 175.4060}$	0.324045 $\boldsymbol{\lambda = 0.27959}$
120	**0.2**	0.8025667 $\boldsymbol{\lambda = 0.98858}$	0.338470 $\boldsymbol{\lambda = 0.02565}$
	0.5	1.290703 $\boldsymbol{\lambda = 1.462254}$	0.4653562 $\boldsymbol{\lambda = 0.028032}$

Sample Size	Correlation Coefficient	MSE Best Penalty parameter	
		RR	Lasso
	0.8	1.576712 $\lambda = \mathbf{1.813312}$	0.6714721 $\lambda = \mathbf{0.617018}$
200	0.2	1.000841 $\lambda = \mathbf{0.95218}$	0.7238574 $\lambda = \mathbf{1.441628}$
	0.5	1.277283 $\lambda = \mathbf{1.338346}$	0.6224570 $\lambda = \mathbf{0.046234}$
	0.8	2.104713 $\lambda = \mathbf{2.04675}$	0.7074679 $\lambda = \mathbf{0.184143}$

Figure 3 shows the fast increase for (MSE) for both (Ridge-Lasso), where the top axis indicates the number of model variables—the number of variables p=60, sample size n=200 with high correlation. We can see in Figure 4 that values of (MSE) in Ridge are increased more than in the Lasso method.



Ridge Method

Lasso Method

Figure 3. The fast increase for (MSE) for both (Ridge-Lasso) where the top axis indicates the number of model variables. Several variables p=60, sample size n=200 with high correlation.

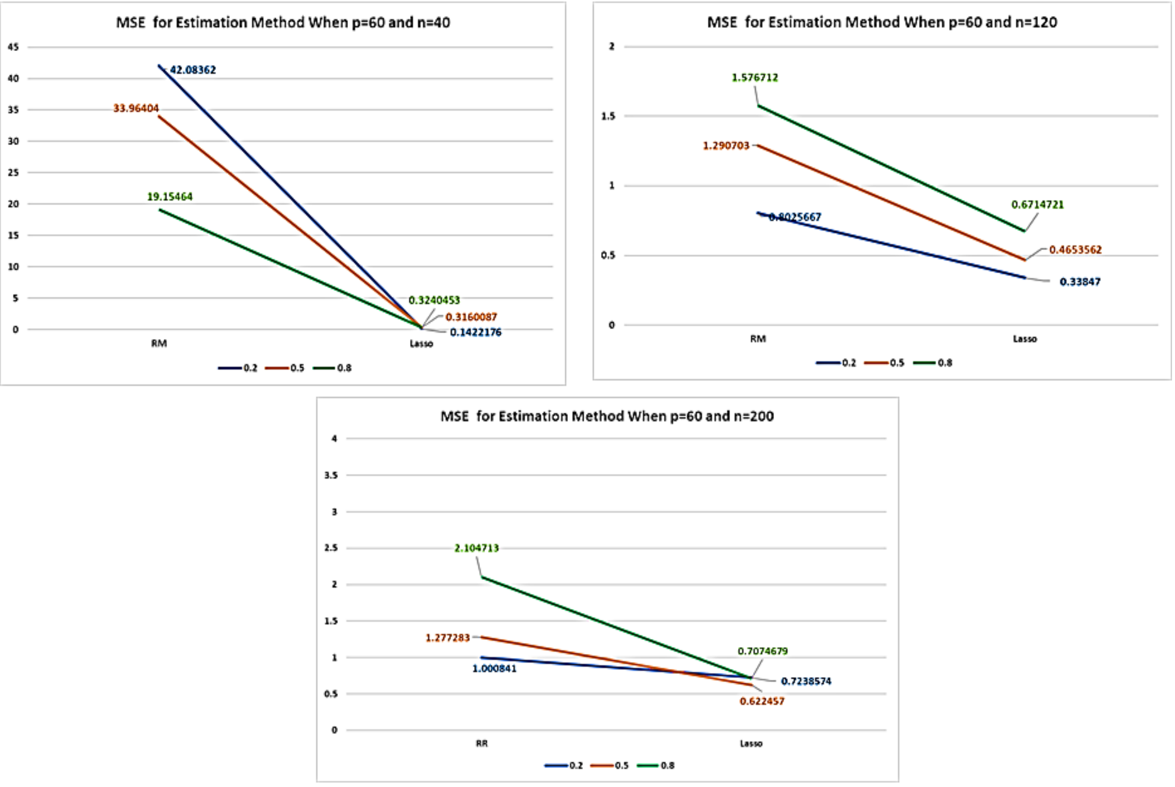


Figure 4. Comparisons between (Lasso-Ridge) via MSE

5. Conclusions

When (p=30, 60) Lasso method has more minor (MSE) than the Ridge method. The Lasso method proved its efficiency by obtaining the least MSE. Optimal Penalty parameter (λ) chosen from Cross-Validation through minimizing (MSE) of prediction. We see a rapid increase for (MSE) for both (Ridge-Lasso) where the Top axis indicates the number of model variables, and when the correlation between variables increases and the sample size too, we can see the (MSE) values increase in the Ridge method, then the Lasso method. A ridge method gives greater efficiency when the sample size is more significant than variables (p<n), but the Ridge method cannot shrink coefficients to precisely zero. So the elasticity of ridge coefficients decreases, but variance increases bias, also (MSE) first remains relatively constant and then increases fast.

Declaration of competing interest

The authors declare that they have no known financial or non-financial competing interests in any material discussed in this paper.

Funding information

No funding was received from any financial organization to conduct this research.

References

[1] A. E. Hoerl, R. W. Kannard, and K. F. Baldwin, “Ridge regression: some simulations,” Communications in Statistics, vol. 4, no. 2, pp. 105–123, 1975.

[2] A. Motlak,"Comparison of Penalized Likelihood Methods for Variable Selection and Parameter Estimation in Poisson Regression " The council of the College of Computer Sciences and Mathematics University of Mosul, 2018.

[3] Y.M. Al-Hassan, M. M.,A. Al-Kassab, "Monte Carlo Comparison between Ridge and Principal Components Regression Methods ", Applied Mathematical Sciences, vol.3, no.42, pp.2085-2098, 2009.

[4] B. M. G. Kibria and A. F. Lukman, "A New Ridge-Type Estimator for the Linear Regression Model: Simulations and Applications", *Scientifica*, vol. 2020, p. 9758378, 2020.

[5] G. Cassela, "Minimax Ridge Regression Estimation ", *The Annals of Statistics*, vol.8,.no.5, pp.1036-1056, 1980.

[6] A. Hazem and A. Yousif, "Comparison between some of the robust penalized estimators using simulation", *Journal of Economics And Administrative Sciences*, vol.32, no.100, pp.490-504, 2017.

[7] S. A. –R. Al-Sabaah and S. M. Al-Kafishi, "Parameters Estimation of the Multiple Linear Regression Model Under Multicollinearity problem", *Journal of economic, administration and financial studies* , vol. 12, no. 1, pp. 1-28, 2020.

[8] F. A. Mahdi, “A comparison of Different Ridge Regression Methods with Application” , MSc Thesis, College of Education for Pure Science / Ibn Al-Haitham, University of Baghdad, 2017.

[9] L. Firinguetti, "A generalized ridge regression estimator and its finite sample properties". *Communications in Statistics-Theory and Methods*, vol. 28, pp.1217-1229, 1999.

[10] M. Goldstein and A.F.M. Smith," Ridge Type Estimation for Regression Analysis", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol.36, pp284-291, 1974.

[11] R.F. Gunst and R.L. Mason, " Biased Estimation Regression: An Evaluation using Mean Squared Error", *Journal of the American Statistical Association*, vol. 72, no.359, pp. 616-628, 1977.

[12] I. A.W. Mohamed, “Effects of Multicollinearity on the Estimation of Macroeconomic Variables by Using Data from Sudan”, Alneelain University, Khartoum, Sudan, p.34, 2012.

[13] L.S. Mayer and T.A. WillK," On Biased Estimation in Linear Models " Techno metric Vol.15 pp497-508, 1974.

[14] M. Ebegil, F. G. Okpınar and M. Ekni,"A simulation study of some shrinkage estimators", *Hacettepe Journal of Mathematics and Statistics*, vol. 35, no. 2, pp.213 – 226, 2006.

[15] M. Irfan, M. Javed and M. A. Raza," Comparison of Shrinkage Regression Methods for Remedy of Multicollinearity Problem", *Middle-East Journal of Scientific Research*, vol.14, no. 4, pp.570-579, 2013.

[16] M. T. Qasaband and M. A. Amin, "Using Ridge Regression Technique for Detecting Multicollinearity with Application" College of Computer Science and Mathematics, University of Mosul, Iraq, 2018.

[17] N. S. Rustum, S. A. Sh. Al-Temimi, (2022)," estimation of the epidemiological model with a system of differential equations (SIRD) using the Runge-Kutta method in Iraq", *International Journal of Nonlinear Analysis and Applications*, vol.13, no. 2, pp.2807-2814, 2022.

[18] A. Norliza, H.A. Maizah, A. Robin, "A comparative study on some methods for handling multicollinearity problems", *Mathematica*, vol.22, pp.109–119, 2006.

[19] O. Toka, "A Comparative Study on Regression Methods in the presence of Multicollinearity", *Journal of Statisticians: Statistics and Actuarial Sciences*, vol.9, no.2, pp.47-53, 2016.

[20] R. Tibshirani," Regression shrinkage and selection via the lasso", Journal *of the Royal Statistical Society: Series B (Methodological)*, vol.58, pp.267-288, 1996.

[21] R. S. Mohammad,"Using Ridge Regression to Study the Effect of some Factors on the General Index of the Stock Market", *Al-Qadisiyah Journal of Administration and Economics Sciences*, vol. 12, no. 4 pp.220-240, 2010.