World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Review Article)

# On using the penalized regression estimators to solve the multicollinearity problem

Mastourah Abdulsatar Ahmeedah [1], Abdelbaset A.Sh Abdalla [2,*] and Ahmed M. Mami [2]

[1] *Department of Statistics, Faculty of Science, University of Ajdabia, Ajdabia, Libya.*

[2] *Department of Statistics, Faculty of Science, University of Benghazi, Benghazi, Libya.*

World Journal of Advanced Research and Reviews, 2024, 24(01), 2654–2664

Publication history: Received on 16 September 2024; revised on 26 October 2024; accepted on 29 October 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.24.1.3265

## Abstract

The paper compares coefficient parameter estimation efficiency using penalized regression approaches. Five estimators are employed: Ridge Regression, LASSO regression, Elastic Net (ENET) Regression, Adaptive Lasso (ALASSO) regression, and Adaptive Elastic Net (AENET) regression methods. The study uses a multiple linear regression model to address multicollinearity issues. The comparison is based on average mean square errors (MSE) using simulated data with varying sizes, numbers of independent variables, and correlation coefficients. The results are expected to be useful and will be applied to real data to determine the best-performing estimator.

**Keywords:** Ridge Regression; LASSO Regression; Elastic Net Regression; Adaptive Lasso Regression; Adaptive Elastic Net Regression methods

## 1. Introduction

The $\boldsymbol{X}$ matrix, which contains the independent variables, causes singularity when some linear combinations of the columns of $\boldsymbol{X}$ are exactly equal to zero. This becomes more evident when the least squares analysis is computed because the unique solution does not exist. The issues that arise from $\boldsymbol{X}$ being nearly singular are known as the multicollinearity problem.

Multicollinearity can lead to inaccurate regression coefficients and difficulty in identifying important variables. Addressing this issue depends on the analysis goals. For prediction, multicollinearity is usually not a significant issue, but when estimating regression coefficients, biased regression methods may be appropriate. Severe multicollinearity can lead to misleading results in identifying important variables. This paper focuses on managing multicollinearity while estimating linear regression model parameters.

## 2. Multicollinearity in Linear Regression Models

Recall that Multicollinearity in linear regression occurs when independent variables have close to zero linear combinations. This can happen due to near-linear dependencies in the data, poorly constructed models, or insufficient sample sizes. Biased Regression is a classic method for addressing multicollinearity.

### 2.1. The Biased Regression Methods

The regression coefficients obtained using the Ordinary Least Squares (OLS) method are known as the Best Linear Unbiased Estimators (BLUE). However, if the multicollinearity problem exists, the minimum variance may be unsatisfactorily large. In the presence of multicollinearity, biased regression methods have been suggested as a potential solution. The best local measure of averaging an estimator to the parameter being estimated is the Mean Squared Error

* Corresponding author: Abdelbaset Abdalla

(MSE).Let $\tilde{\theta}$ be a biased estimator having a smaller Mean Squared Error (MSE) than an unbiased estimator $\theta$, the Mean Squared Error of $\tilde{\theta}$ can be defined as

$$\text{MSE}(\tilde{\theta}) = E(\tilde{\theta} - \theta)^2 \text{ ................ (1)}$$

Remember that the variance of an estimator $\tilde{\theta}$ can also be defined as

$$\text{Var}(\tilde{\theta}) = E\left[\tilde{\theta} - E(\tilde{\theta})\right]^2 \text{ ................ (2)}$$

In expression (1), the MSE of $\tilde{\theta}$ is calculating the average squared deviation of the estimator from the parameter being estimated, whereas in expression (2), the variance of $\tilde{\theta}$ is calculating the average squared deviation of the estimator from its expectation.

If the estimator is unbiased, then $E(\tilde{\theta}) = \theta$ and MSE($\tilde{\theta}$) =$\sigma^2(\tilde{\theta})$. If the estimator is biased, then the MSE is equal to the variance of the estimator plus the square of its bias, where the Bias ($\tilde{\theta}$) = $E(\tilde{\theta}) - \theta$. It is very possible for the biased estimator to obtain a variance that is sufficiently smaller than the variance of an unbiased estimator in order to compensate for the bias introduced.

The biased regression technique is based on the idea that, on average, the biased estimator will be closer to the true parameter, even though its mean will not be equal to the true parameter. This trade-off between bias and variance is illustrated in Figure 1.

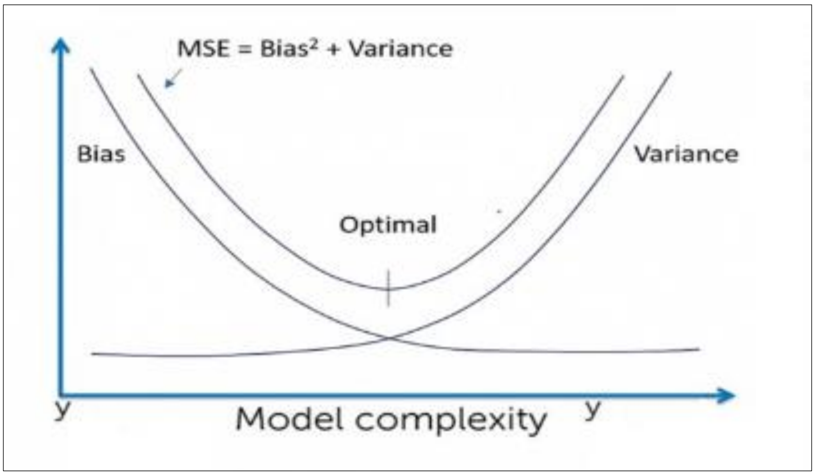



**Figure 1** Illustration of the process of Estimation in Biased Regression

The possible advantage of biased estimators is shown in Figure 1. Therefore, it may be possible to find an estimator for which the sum of its squared bias and its variance (i.e. the MSE) is smaller than the variance of the unbiased estimator [7].

## 3. Review the Development of Bias Regression Methods

It's important to use biased regression methods cautiously to address multicollinearity. Ridge regression and principal component regression are commonly used for this purpose. Penalized regression approaches, such as Ridge regression, LASSO regression, Elastic Net (ENET) Regression, Adaptive Lasso (ALASSO) regression, and Adaptive Elastic Net (AENET) regression methods, are employed to compare coefficient parameter estimation efficiency using simulated research data with small sample sizes and varying numbers of independent variables.

### 3.1. The Ridge Regression

**Ridge** Regression is commonly used to address multicollinearity in Least Squares estimation. The Ridge Regression estimator is developed by examining the Mean Squared Error (MSE) of the least squares estimator of $\beta$

$$\text{MSE}(\hat{\boldsymbol{\beta}}) = \text{E}||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2 \text{ .................... (3)}$$

and it can be rewritten in the following form:

$$E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = \sum_j E(b_j - \beta_j)^2 = \sum_j \{E(b_j) - \beta_j\}^2 + \sum_j Var(b_j) \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

It is widely recognized that the least squares method achieves the smallest variance among all unbiased linear estimates, as stated by the Gauss-Markov theorem. However, it's important to note that the minimum mean squared error (MSE) is not always guaranteed. To gain a better understanding of this concept, it's helpful to explore the various types of penalized estimation methods. Regression estimators, let $\hat{\boldsymbol{\beta}}^{LS}$ denote the ordinary least squares estimator of $\boldsymbol{\beta}$.

The multiple linear regression model can be seen as,

$$Y_{(N\times 1)} = X_{(N\times(P+1))}\beta_{((P+1)\times 1)} + \varepsilon_{(N\times 1)}$$

the estimator $\hat{\beta}^{LS} = (X^tX)^{-1}X^ty$ is unbiased estimator of $\beta$, also

$$E(\hat{\beta}^{LS}) = \beta \text{ and } Cov(\hat{\beta}^{LS}) = \sigma^2.(X^tX)^{-1}$$

$$\text{Hence, } MSE(\hat{\beta}^{LS}) = E\|\hat{\beta}^{LS}\|^2 - \|\beta\|^2.$$

$$= tr\{\sigma^2(X^tX)^{-1}\} = \sigma^2.tr\{(X^tX)^{-1}\} \ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Therefore, by rearrange (5), we obtain

$$E\left(\|\hat{\beta}^{LS}\|^2\right) = \|\beta\|^2 + \sigma^2.tr\{(X^tX)^{-1}\} \ldots\ldots\ldots\ldots\ldots\ldots (6)$$

Because the unique solution of $X^tX$ does not exist, the resultant least square estimate of $\hat{\beta}^{LS}$ would be large in length $\|\hat{\beta}^{LS}\|$ and it is related to large standard errors. This large variation would also lead to poor model prediction.

The **Ridge** Regression is a constrained type of least squares. It solves the estimation problem by producing a biased estimator, however, with small variances [10].

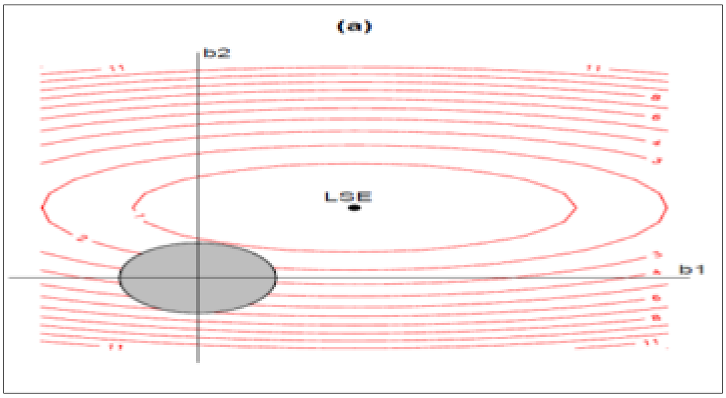



**Figure 2** Contours of the Sum of Squares of the Residual and the $\boldsymbol{L_2}$- Constraint Functions in Ridge Regression

From a Lagrangian problem point of view, it is equivalent to minimizing

$$Q^*(\beta) = \|\beta\|^2 + (1/k)\{(\beta - \hat{\beta}^{LS})^t X^t X(\beta - \hat{\beta}^{LS}) - \phi_0\} \ldots\ldots\ldots\ldots (7)$$

where $k$ is considered to be a deflection factor selected to satisfy the constraint.

Therefore, using the differentiation of $Q^*(\beta)$with respect to $\beta$

$$\frac{\partial Q^*(\beta)}{\partial \beta} = 2\beta + (1/_k)\{2(X^tX)\beta - 2(X^tX)\hat{\beta}^{LS}\} = 0 \text{ ................(8)}$$

that derives the **Ridge** estimator as follows:

$$\hat{\beta}^R = \{X^tX + k\text{I}\}^{-1}X^ty \text{ ...................(9)}$$

When using Ridge Regression, selecting the best value for $k$ is crucial. The practical option is cross-validation, which finds the optimal value for maximum prediction accuracy [5]. The cross-validation approach for choosing $k$ is as follows:

- On the training set: estimate several different Ridge Regression models with different values of $k$.

- On the validation set: choose the best model (best $k$ which gives the lowest MSE on the validation set).

### 3.2. The Least Absolute Shrinkage and Selection Operator (LASSO)

The **LASSO** is a type of Panelized Regression method, similar to Ridge Regression but with the important feature of variable selection. While Ridge Regression makes the selection process continuous by adjusting the reduction parameter, the LASSO sets some coefficients to zero, aiming to combine the advantages of both subset selection and Ridge regression. The LASSO estimator of $\beta$ is obtained by

minimizing $\|y - \text{X}\beta\|^2$, subject to $\sum_{j=1}^{p} |\beta_j| \leq s$.

$$\hat{\beta}(\text{lasso}) = \arg\min_{\beta} \left( \frac{\|\mathbf{y} - \mathbf{X}\beta\|_2^2}{n} + \lambda \|\beta\|_1 \right),$$

More explicitly, the $L_2$ penalty $\sum_J \beta_j^2$ in **Ridge** Regression is substituted by the $L_1$ penalty $\sum_j |\beta_j|$ in **LASSO**. If one chooses $s \geq \sum_j |\beta_j^{LS}|$, then the **LASSO** estimates are the same as the Least Squares estimation. If one chooses $s \leq \sum_j |\beta_j^{LS}|$, then it will cause reduce of the solutions towards zero.

Figure 2.2 exemplifies the contours of the residual sum of squares together with the $L_1$ **LASSO** constraint in the two-dimensional case [2].

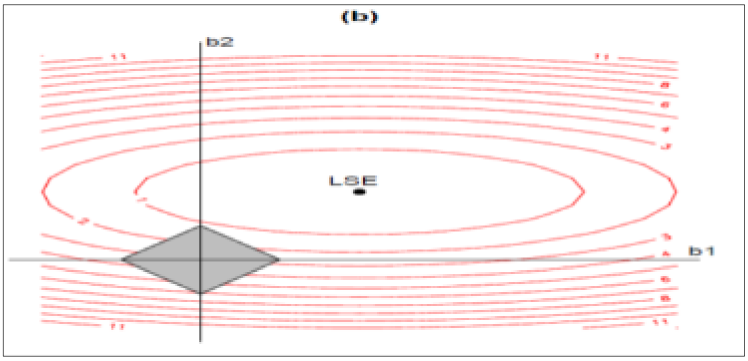



**Figure 3** Contours of the Sum of Squares of the Residual and the $\boldsymbol{L_1}$-Constraint Functions in LASSO Regression

In Figure 2, the constraint region in Ridge Regression is disk-shaped, while in Figure 3, the constraint region in LASSO is diamond-shaped. Both methods start by finding the first point where the elliptical contours hit the constraint region. If the solution occurs at a corner in LASSO, then one coefficient is equal to zero. Moreover, the LASSO solution is quite similar to the Ridge Regression solution but with many zero coefficient estimates. In the case of orthonormal designs where $X^tX = I$, the LASSO estimator can be written as

$$\hat{\beta}_j^{lasso} = \text{sign}\,(\hat{\beta}_j^{LS})\{|\hat{\beta}_j^{LS}| - \gamma\}_+ \text{ ................ (10)}$$

Where $\gamma$ is constrained by the condition $\sum_j |\hat{\beta}_j^{lasso}| = s$

In summary, the coefficients whose values are greater than the threshold $\gamma$ would be contracted by a unit of $\gamma$, whereas the coefficients whose values are smaller than $\gamma$ would be automatically forced to go to 0. Therefore, the **LASSO** procedure performs as a variable selection operator [1].

[9] in his first work on LASSO used quadratic programming to solve the optimization problem, because of the non-smooth performance of the LASSO constraint. It is based on the fact that the condition $\sum_j |\beta_j| \leq s$ is equivalent to $\delta_i^t$ for all $i$= 1, 2, . . . , $2^p$, where $\delta_i$is the $p$-tuples of form (±1,±1, . . . ,±1).

While [6] developed a compact descent method to solve the constrained LASSO problem for any fixed $s$.

The paper explores how coefficients change as the parameter λ varies. Ridge Regression reduces coefficients together as λ increases, while LASSO reduces some coefficients to zero before others as λ increases. Ridge Regression tends to reduce coefficients uniformly as λ increases, while LASSO reduces coefficients unevenly, allowing some to reach zero before others as $\lambda$ increases.

### 3.3. The Adaptive LASSO Estimator

Let us consider the weighted **LASSO** Regression as

$$\arg\min_{\beta} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j|, \qquad \text{................ (11)}$$

where $w_j$ is a known weights vector. If the weights are data dependent and carefully selected, then the weighted **LASSO** can have the Oracle Properties.

The new methodology is called the **Adaptive LASSO** (**ALASSO**) Regression.

### 3.4. The Elastic Net (ENET) Regression

combines **LASSO** and **Ridge** Regression to address multicollinearity and overfitting in high-dimensional datasets. It adds penalty terms to the least squares objective function and uses $L_1$ and $L_2$ norms for feature selection and reduction. Introduced by [12], this linear regression algorithm is a powerful tool in machine learning. The Elastic Net (**ENET**) Regression model can be represented as follows:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots\ldots\ldots + b_n x_n + e$$

where y is the dependent variable, $b_0$ is the intercept, $b_1,\ldots,b_n$ are the regression coefficients, $x_1,\ldots,x_n$ are the independent variables, and e is the error term. The Elastic Net (**ENET**) Regression model tries to minimize the following objective function :

$$[\ \mathrm{RSS} + \lambda\ [(\ 1\text{-}\propto) * \|\ \beta\ \|^2 + \alpha\ \|\ \beta\ \|^1]$$

where RSS is the residual sum of squares, the regularization parameter, $\beta$, is the coefficient vector, α is the mixing parameter between the $L_1$ and $L_2$, the norms, ‖ $\beta$ ‖$^2$ is the $L_2$ norm of $\beta$ and ‖ $\beta$ ‖$^1$ is the $L_1$ norm of $\beta$.

Given dataset ($y$ , $x$), and define an artificial data (y*, x*) by

$$\mathbf{x}^*_{(n+p)\,xp} = \frac{1}{\sqrt{1+\lambda_2}} \left( \frac{x}{\sqrt{\lambda_2\, I}} \right), \ \mathbf{y}^*_{(n+p)} = \begin{pmatrix} Y \\ 0 \end{pmatrix}.$$

Both [12] and [8] have found that the Elastic Net (ENET) estimator performs very well when compared with the Ridge and LASSO regression. The Elastic Net (ENET) estimates $\beta_{ENET}$ can be derived as follows.

Using the definition of Ridge and LASSO regression, yields

$$\hat{\beta}_{ENET} = \arg \underset{\beta\ \ ER^P}{min} \left( \| y^* - x^* \frac{\beta}{\sqrt{1+\lambda_2}} \| + \frac{\lambda_1}{\sqrt{1+\lambda_2}} \| \frac{\beta 2}{\sqrt{1+\lambda_2}} \|_1 \right). \quad 2$$

$$= \arg \min_{\beta \in R^P} \beta^T \left(\frac{X^{*T}X^*}{1+\lambda_2}\right)\beta - 2\frac{Y^{*T}X^*}{\sqrt{1+\lambda_2}}\beta + Y^{*T}Y^* + \frac{\lambda_1 \| \beta \|_1}{1+\lambda_2}. \quad \text{.........} (12)$$

$$\text{where } X^{*T}X^* = \left(\frac{X^TX+\lambda_2}{1+\lambda_2}\right),\ Y^{*T}X^* = \frac{Y^TX}{\sqrt{1+\lambda_2}},\ \text{and } Y^{*T}Y^* = Y^TY$$

By substituting into equation (11) yields the following

$$\hat{\beta}(\text{enet}) = \left(1+\frac{\lambda_2}{n}\right)\left\{\arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_2\|\beta\|_2^2 + \lambda_1\|\beta\|_1\right\}$$

In this case, the minimizing the objective function Q of the Elastic Net (**ENET**) is given by

$$Q_{ENET}(\beta, \lambda_2, \lambda_1) = (1+\lambda_2)\left[OLS(\beta) + \lambda_2 P_2(\beta) + \lambda_1 P_1(\beta)\right]$$

$$= (1+\lambda_2)\left[\| y - x\beta \|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1\right]$$

$$= \beta^T\left(\frac{X^TX+\lambda_2 I}{1+\lambda_2}\right)\beta - 2Y^TX\beta + \lambda_1 \| \beta \|_1 \quad \text{.....................} (13)$$

$$= (1+\lambda_2)\left[\sum_{i=1}^{n} \left(y_i - \beta_o - \sum_{j=1}^{p} B_J X ij\right)^2 + \lambda 2 \sum_{j=1}^{p} B_j^{\ 2} + \lambda 1 \sum_{j=1}^{p} | \beta | B_J\right]$$

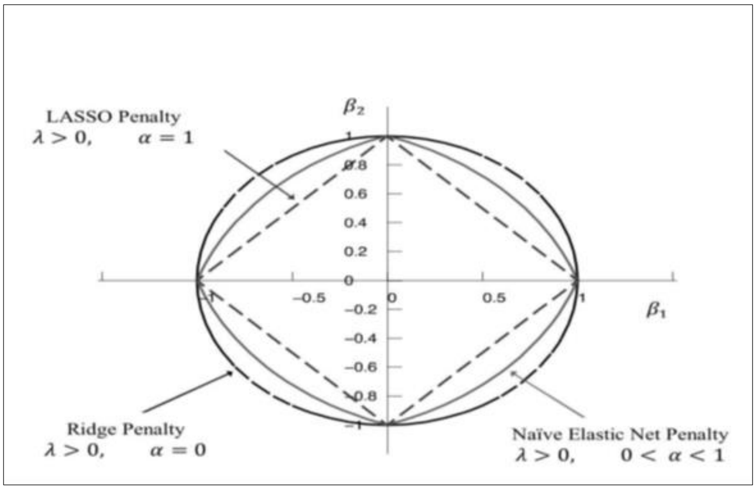



**Figure 4** Contour Plots for the Ridge, LASSO and Elastic Net Regression estimators

Figure 4 illustrates the suitability of Ridge, LASSO, and Elastic Net (ENET) Regression in a two-dimensional setting. The Elastic Net estimate requires an iterative algorithm, and a popular algorithm for this is the Least Angle Regression (**LARS**). The Elastic Net Regression works by adding a penalty equivalent to the sum of the absolute values ($L_1$-norm) of the coefficients and the squares ($L_2$-norm) of the coefficients.

### 3.5. The Adaptive Elastic Net Estimator

Elastic Net Regression combines $L_1$ and $L_2$ penalties to reduce coefficients and set some to zero. The penalty amount can be adjusted using constants ($\lambda_2$ and $\lambda_1$). Adaptive Elastic Net (**AENET**) regression minimizes the Penalized Least Squares (PLS) objective function.$PLS_{AENET}(\beta, \lambda_1, \lambda_2) = OLS(\beta) + \lambda_2 P_2(\beta) + \lambda_1 P_1(\beta)$

$$= || y - X\beta||_2^2 + \lambda_2|| \beta||_2^2 + \lambda_1 ||\beta||_1$$

or

$$= \sum_{i=1}^{n}(yi - \beta o - \sum_{j=1}^{p} \beta i x i j)^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 + \lambda \sum_{j=1}^{p} ||\beta_j||$$

$$\hat{B}_{AENET} = \arg\min_{B \in R^p} [||y - X\beta||_2^2 + \lambda_2||\beta||_2^2 + \lambda_1||\beta||_1] \quad (14)$$

[9] has observed that the empirical predictive performance of **LASSO** is dominated by **Ridge** regression (i.e., $L_2$-norm regularized linear regression) when the independent variables are highly correlated, while the situation is reversed when there is a relatively small number of more independent variables.

Therefore, combining both $L_1$-norm and $L_2$-norm may be necessary to achieve "the best of both the **Ridge** and **LASSO** Regression Features.

$$\text{Let} \quad \lambda_2 = \frac{1}{2}\lambda(1-\alpha) \text{ and } \lambda_1 = \lambda_\alpha, \ldots\ldots\ldots\ldots\ldots\ldots (15)$$

the Penalized Least Squares (PLS) or the objective function of the Adaptive Elastic Net (**AENET**) using equation (15) becomes as follows:

$$PLS_{AENET}(\beta,\lambda,\alpha) = || Y - X\beta ||_2^2 + \lambda(\frac{1}{2}(1-\alpha)||\beta||_2^2 + \alpha\, ||\beta||_1)$$

$$= \sum_{i=1}^{n}\left(y_i - \beta_o \sum_{j=1}^{p}\beta_j\, x_{ij}\right)^2 + \lambda\left(\sum_{j=1}^{p}\left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha\,|\beta|\right]\right) \ldots\ldots\ldots\ldots(16)$$

where $\lambda > 0$ is the penalty (or the tuning) parameter of the Adaptive Elastic Net (**AENET**) regression and $\alpha$ is the mixing parameter and controls the influence of $L_1$ and $L_2$ penalizes [11].

The *glmnet* library in R Package (version 4.1.0) handles Elastic Net (**ENET**) regression, prov**id**ing a flexible framework through **Ridge** and **LASSO** regularized procedures. *It* includes the "aenet" function for Adaptive Elastic Net (**AENET**) regression, **which** enforces sparsity and a grouping effect and performs best when close to **Ridge** regression or **LASSO** Regression. The **AENET** estimator involves a two-stage reduction procedure for **Ridge** and **LASSO** regression coefficients.

## 4. Simulation Study

In this section, we compare the performance of five estimators: Ridge, Lasso, Elastic Net, Adaptive Lasso, and Adaptive Elastic Net, used to address multicollinearity issue through simulation.

In this simulation study, we are comparing five penalized regression estimators by varying the correlation coefficients between predictors. The goal is to determine the most effective estimator for strong correlations. We are considering two scenarios: Pattern 1 with normal distribution error terms and Pattern 2 with non-normal distribution error terms. In each Pattern, we explore different numbers of independent variables (p = 2, 5, and 10) and sample sizes of n = 50, 100, and 300. The model is represented as $Y = f(X) + \varepsilon$.

Finally, we incorporate two distinct error distributions: the normal distribution and the heavy-tailed $t$ distribution.

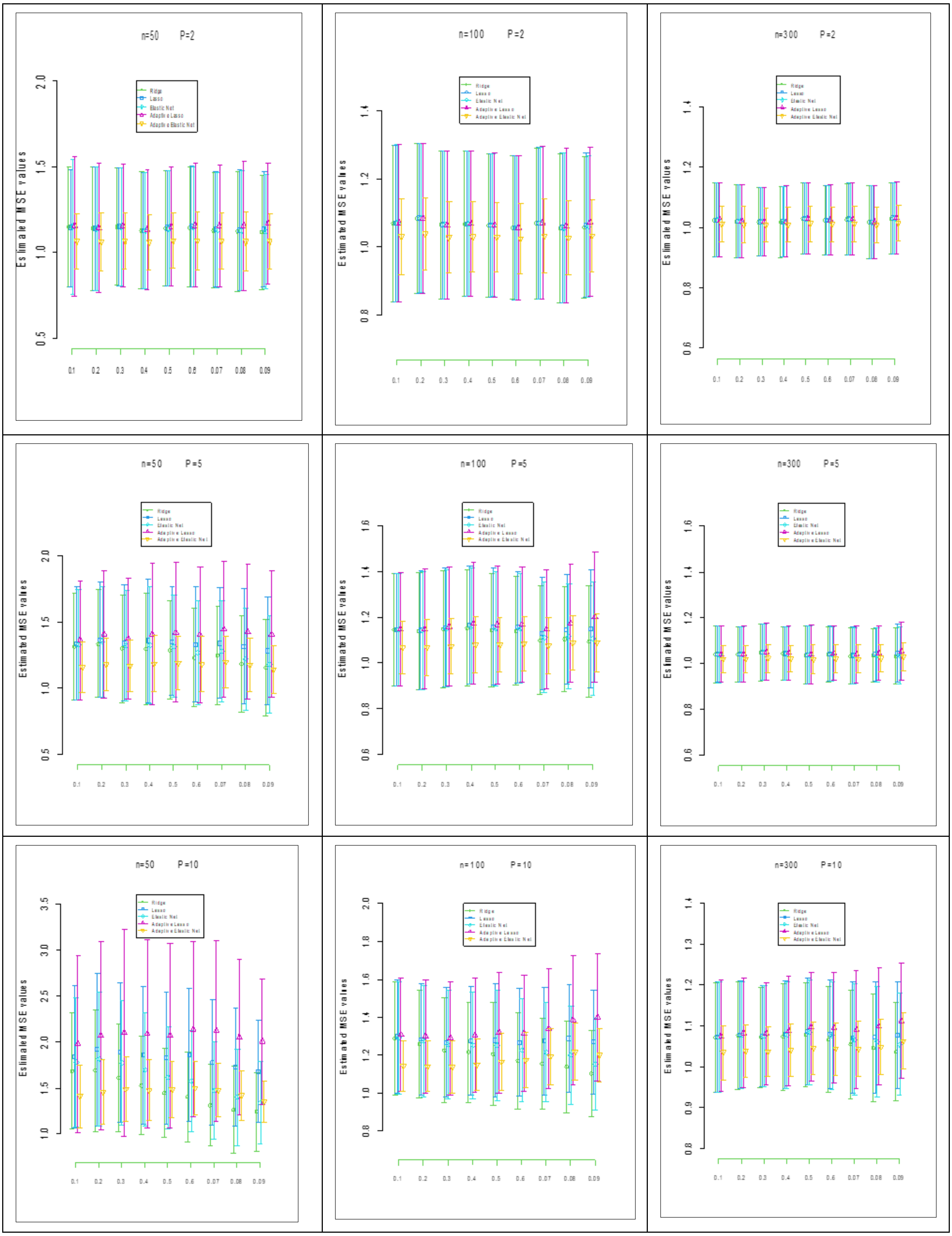
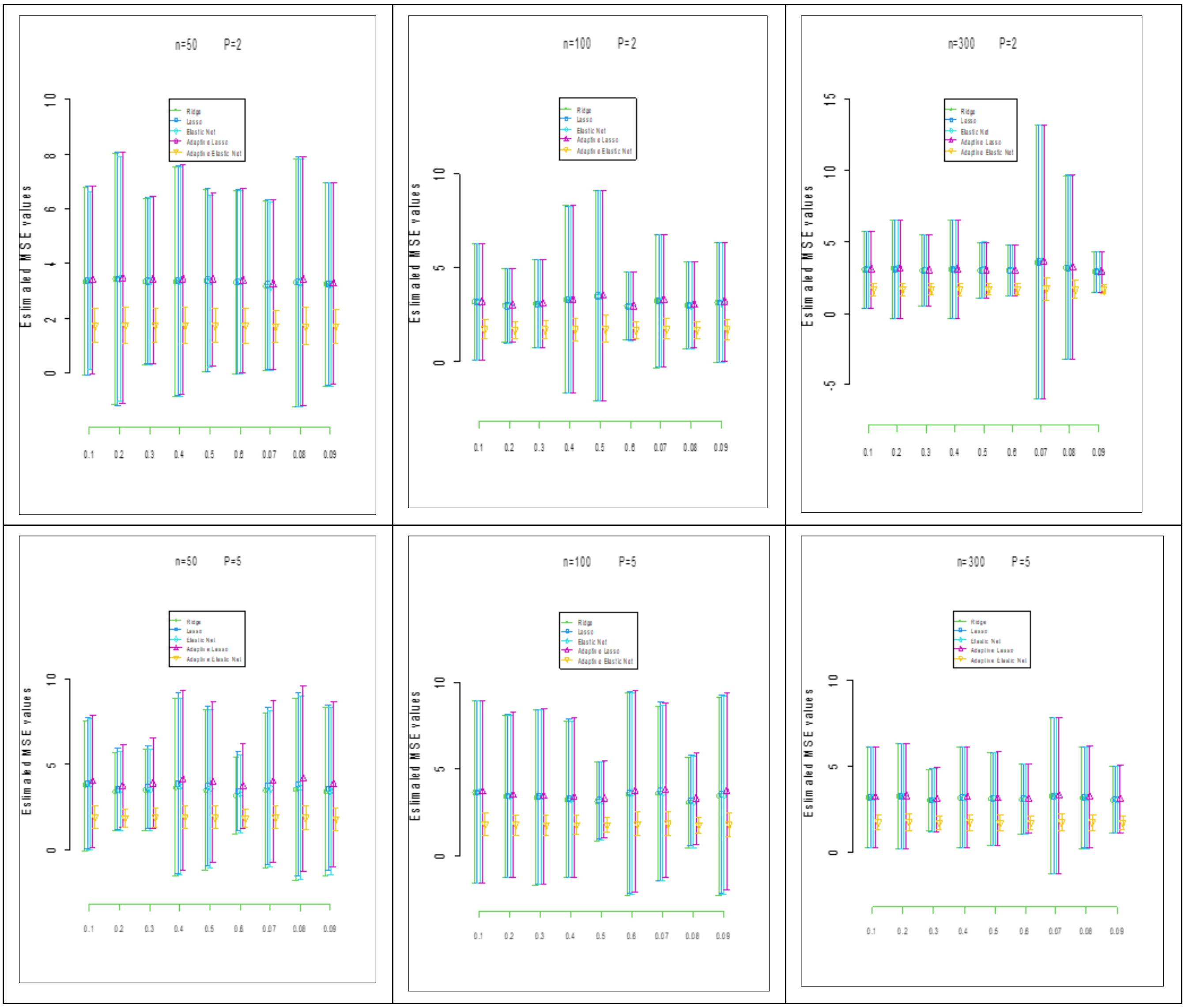



**Figure 5** The Box plots demonstrate how various choices of correlation ρ affect the mean (standard deviation) of the MSE value for the five proposed estimators in Pattern 1

### 4.1. Pattern 1 results, where error terms follow a normal distribution

Figure 1 shows how different correlation values (ρ) affect the mean and standard deviation of the Mean Squared Error (MSE) for the five proposed estimators in Pattern 1. Analysis based on the MSE criterion indicates that the Adaptive Elastic Net consistently outperforms the other estimators across various predictor configurations and sample sizes. For instance, when p=2 and n=50, the Adaptive Elastic Net ranks first, followed by Ridge, and at certain correlation values, Elastic Net takes second place. For n=100 and n=300, the Adaptive Elastic Net remains the top performer, with Lasso and Ridge alternating in second place depending on the correlation parameter ρ. Similarly, for p=5 and p=10, regardless of sample size, the Adaptive Elastic Net consistently leads, followed by Ridge and then Elastic Net. It's worth noting that all estimators show small standard deviations, but the Adaptive Elastic Net consistently maintains the smallest standard deviation across all scenarios, indicating its robustness in estimation accuracy and variable selection performance.



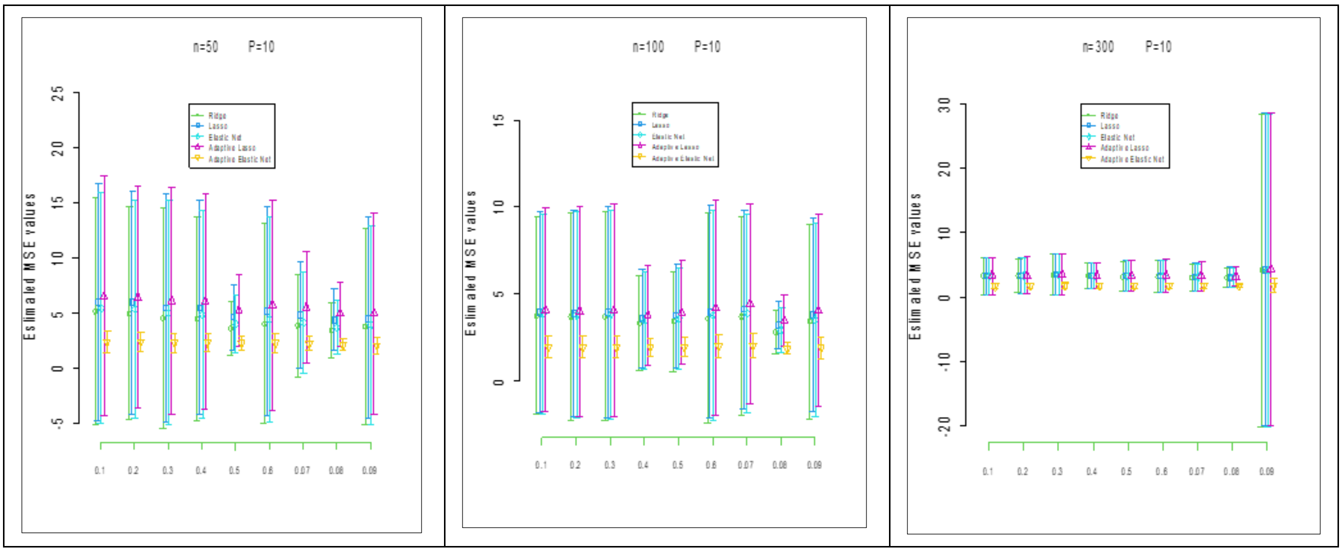



**Figure 6** The Box plots demonstrate how various choices of correlation ρ affect the mean (standard deviation) of the MSE value for the five proposed estimators in Pattern 2

### 4.2. Pattern 2 results, where error terms follow a non-normal distribution

The analysis of Figure 6 reveals the performance of various estimators based on the Mean Squared Error (MSE) criterion across different configurations of predictors and sample sizes. In the case of p = 2 and n = 50, the Adaptive Elastic Net consistently outperforms all other estimators regardless of the correlation parameter ρ, with the OLS estimator performing poorly. The Elastic Net ranks second for certain values of ρ, while Ridge takes this position for others. For n = 100 and n = 300, the Adaptive Elastic Net remains the top performer, with Lasso and Ridge alternating in second and third places depending on ρ. In scenarios with p = 5 $ and p = 10, the Adaptive Elastic Net again excels, followed by Ridge in second and Elastic Net in third, regardless of sample size. Notably, throughout all configurations, the Adaptive Elastic Net exhibits the smallest standard deviations, highlighting its robustness in estimation accuracy and variable selection performance.

## 5. Conclusion

This section evaluates five penalized regression estimators for addressing multicollinearity, focusing on Ridge, LASSO, Elastic Net, Adaptive LASSO, and Adaptive Elastic Net. Ridge regression mitigates multicollinearity by introducing bias to reduce variance, while LASSO and Elastic Net aim to eliminate variables most affected by multicollinearity. In the first pattern with normally distributed error terms, OLS consistently yields high MSE values due to multicollinearity. The Adaptive Elastic Net emerges as the top performer across various sample sizes and correlation coefficients, followed by Ridge. In the second pattern with non-normally distributed errors, OLS again performs poorly, while the Adaptive Elastic Net remains superior, closely followed by Ridge and LASSO in certain cases. Notably, regardless of the error distribution, the Adaptive Elastic Net consistently ranks first, demonstrating robustness against non-normality. This estimator effectively reduces the number of selected variables while maintaining accuracy, making it preferable for real-world variable selection challenges. Overall, it outperforms other penalized estimators by achieving the smallest MSE and standard deviation values, while Ridge and LASSO follow closely behind.

## Compliance with ethical standards

*Disclosure of conflict of interest*

No conflict of interest to be disclosed.

## References

[1] Baltagi, B. H. (2001). Econometric Analysis of Panel Data. Wiley, John & Sons:New York.

[2] Greene, W. H. (2000). Econometric Analysis, 4th Edition. Prentice.

[3] Helwig, N. E. (2017). Adding bias to reduce variance in psychology results: A tutorial on Penalized regression. The Quantitive Methods for Psychology, 13(1) : 1-19

[4] Hoerl A.E., Kennard R.W. (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. Vol. 12: pp 55–67.

[5] Kotz, S. and Nadarajah, S. (2004). Multivariate t Distributions and Their Applications. Cambridge University Press.

[6] Osborne, J. W., et al. (2000). A New Approach to Multicollinearity in Multiple Regression: The LASSO.

[7] Park, S. H. (1981) " Collinearity and Optimal Restrictions on Regression Parameters for Estimating Responses" Technometrics 23(1) pp: 289-295

[8] Rish, I. and Grabarnik, R. D. (2015). Sparse Modelling, Algorithms and Applications. First Edition, CRC Press, Taylor and Francis Group, USA.

[9] Tibshirani R. (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodol). 1;267–288.

[10] Weisberg, S. (2005). Applied Linear Regression. 3rd edition. Wiley and Sons, Inc.

[11] Xiao, Nan and Xu, Qing-Song (2015) Multi-step adaptive elastic-net: reducing false positives in high-dimensional variable selection. Journal of Statistical Computation and Simulation, pp 1-11

[12] Zou H. and Hastie T. (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statist Methodol);67:301–320.