Estimating the Linear Regression Model in High-Dimensional Data and collinearity

Sahar Hassan, Nahed Helmy and

Amira Elbadawy

Department of statistics, Faculty of Commerce (Girls' Branch), Al-Azhar University, Cairo, Egypt

Estimating the Linear Regression Model in High-Dimensional Data and collinearity

Sahar Hassan, Nahed Helmy and

Amira Elbadawy

Department of statistics, Faculty of Commerce (Girls' Branch), Al-Azhar University, Cairo, Egypt

Abstract

This paper is concerned with introducing the most used penalized regression methods, including ridge regression (RR), least absolute shrinkage and selection operator (LASSO), and elastic net (EN) regression for estimating the linear regression model. These models are used in two cases low and high-dimensional data when data is contain outliers when the explanatory variables have collinearity among them. The Monte Carlo simulation study is conducted to evaluate and compare the performance of these estimators. The simulation results indicate that the obtained estimators using EN are efficient and reliable than the other estimators.

Keywords: penalized regression; Ridge regression; least absolute shrinkage and selection operator; Elastic net; High-dimensional data; Collinearity; Outliers.

1. Introduction

Regression analysis is the most useful statistical technique for analyzing multifaceted data in numerous fields such as science, engineering, and social sciences. Regression analysis is used to study the relationship between the response variable Y_i and one or more explanatory variables X_{ij} which is called linear regression model [Adegoke, (2016)]. The case of one explanatory variable is called simple linear regression model while the case with two or more explanatory variables is called multiple linear regression model. The assumptions of the linear regression model are:

1. The response variable Y is normally distributed as

$$Y_i \sim N(\mu_i, \sigma^2), \qquad i = 1, \dots, n$$

with mean μ_i and variance σ^2.

2. The linearity of the model, (i.e., a linear relationship between the response variable and the explanatory variables).

The general form of multiple linear regression model is formulated as follows:

$$Y = X\beta + \epsilon \qquad (1)$$

where Y is an (n×1) vector of response variable, X is an (n×k) design matrix of explanatory variables, β is a (k×1) vector of unknown parameters and Є is an (n×1) vector of random errors with mean zero and fixed variance (σ^2)

$$[\ \epsilon \sim N(0, \sigma^2 I)\ ]$$

The OLS estimator is given by:

$$\hat{\beta} = (\acute{X}X)^{-1}\acute{X}Y \quad , \qquad (2)$$

Where $\acute{X}$ is the transpose of the matrix X.

The goal of linear regression is to fit a straight line to few points minimizing the sum of squared residuals. Regression models are used for many purposes, such as analysis of variance (ANOVA), estimating the parameters, prediction and selection of the variables.

The ordinary least square (OLS) method is the best linear unbiased estimators (BLUE). But data analysts pointed out some deficiencies of the OLS with respect to prediction accuracy and interpretation, the OLS may not exist if the design matrix is singular [Neter *et al*. (2005)].

Hoerl and Kennard (1962) introduced the RR estimators as an alternative to OLS estimators in the presence of multicollinearity. It leads to estimators which have mean square error (MSE) smaller than the estimators of OLS.

Tibshirani (1996) proposed LASSO to overcome the deficiencies of OLS such as prediction and interpretation of the reduced model. It is a powerful method that perform two main tasks: regularization and selection of variables.

Zou and Hastie (2005) proposed a new regularization and selection of variables method, it is called EN. They viewed the EN as a generalization of the LASSO which overcomes the limitations of the second one. This method is very useful when $k > n$ or there are many correlated variables.

Tibshirani (1997) proposed a new method for selection of variables and shrinkage in Cox's proportional hazards

model. This method is diffident of the LASSO proposed by Tibshirani, designed for the linear regression context.

Osborne *et al.* (2000) proposed a new approach to selection of variables in least squares problems.

Fan and Li (2001) proposed a selection of variables method via penalized likelihood approaches. The algorithm gives estimators with good statistical properties.

Hui and Zou (2006) proposed a new version of the LASSO called the adaptive LASSO. for simultaneous estimation and selection of variables, where adaptive weights are used for penalizing different coefficients in the ℓ_1 penalty.

Hebiri and Van (2011) introduced the LASSO-Type estimator $\hat{\beta}^{Quad}$which consists of two penalty terms: a ℓ_1 penalty term which ensures sparsity and a quadratic penalty term which captures some structure in the regression vector. This estimator satisfies good theoretical properties, specifically when the LASSO estimator might fail.

Arslan (2012) introduced the weighted least absolute deviation WLAD-LASSO method to improve the robustness of the OLS and least–sum of absolute deviations (LAD) based on LASSO method.

Fonti and Belitser (2017) explained and discussed the use of the LASSO method to select the explanatory variables when applying the LASSO to a linear regression model, generalized linear model and logistic regression model for a high-dimensional dataset.

Melkumova and Shatskikh (2017) introduced the comparison of RR and LASSO estimators. All the required

calculations are performed using the R software for statistical computations.

Huang *et al.* (2018) proposed a constructive approach for estimating sparse, high-dimensional linear regression models.

Emmert-Streib and Dehmer (2019) introduced high-dimensional LASSO-Based computational regression models: regularization, shrinkage, and selection.

Weigeet*et al.* (2019) applied LASSO regression method to analyze the influencing factors of vegetable price and more accurate results have been achieved which are based on cucumber price data and influencing factor data.

Januaviani*et al.* (2019) introduced the LASSO method to predict Indonesian foreign exchange deposit data.

This paper is organized as follows. The penalized regression methods are discussed in Section (2). Section (3) is devoted to the Monte Carlo simulation study. Some concluding remarks are presented in Section (4).

2. Penalized Regression Methods

Penalized regression methods keep all the predictor variables in the model but constrain (regularize) the regression coefficients by shrinking them toward zero. If the amount of shrinkage is large enough, these methods can also perform as selection of variables by shrinking some coefficients to zero. These methods are formulated in the constrained minimization form, where the solution for the vector of regression coefficients, is obtained by minimizing the sum squares of regression (SSR) subject to a penalty on the regression coefficients. The shrinkage (tuning) parameter

t determines the amount of shrinkage on the regression coefficients. In the last decade, many different penalized regression methods have been proposed. The LASSO method (Tibshirani1996), adaptive LASSO (Zou 2006) and Elastic Net (Zou and Hastie 2005) are the most popular. For each method, the penalty t imposed on the regression coefficients takes a different form [Hui and Zou, (2006)].

2.1Ridge Regression

RR method is to remedy multicollinearity problems by modifying the method of OLS to allow biased estimators of the regression coefficients. The advantage of a RR method can be reduce the variance by paying the price of an increasing bias. This can be improve the prediction of accuracy of a model. This works is appropriate in situations where the OLS estimators have a high variance $k < n$. A disadvantage of the RR is that it does not shrink coefficients to zero and, does not taking the select the variables.

The RR estimator can be viewed as an OLS estimator with an additional penalty imposed on the coefficient vector.

The general form of the minimum of the penalized residual sum squares is given as :

$$\min \|Y - X\beta\|_2^2$$

$$\text{subject to } \|\beta\|_2^2 \leq t. \qquad (3)$$

The model can be formulated as follows:

$$\hat{\beta}^{RR}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left(\frac{\|Y - X\beta\|_2^2}{2n} + \lambda \|\beta\|_2^2 \right), \qquad (4)$$

where $\|Y - X\beta\|_2^2 = \sum_{i=0}^{n} (Y_i - (X\beta)_i)^2$, $\hat{\beta}_{RR}$ is the vector of the standardized ridge regression coefficients, $\|\beta\|_2^2 = \sum_{j=1}^{k} \beta_j^2$,

and $\lambda > 0$ is called the regularization parameter, is the tuning or regularization parameter that controls the shrinkage of coefficients [Emmert-Streib and Dehmer, (2019)].

2.2 LASSO Regression

LASSO is a regularization and selection of variables method for statistical models which minimizes the sum of squared errors, with an upper bound on the sum of the absolute values of the model parameters. The estimator is defined by the solution for the ℓ_0 optimization problem

$$\text{Minimize}\left(\frac{\|Y-X\beta\|_2^2}{n}\right)$$

$$\text{subject to}\sum_{j=1}^{k}\|\beta\|_1 < t, \qquad (5)$$

where t is the upper bound for the sum of the coefficients. This optimization problem is equivalent to the parameter estimator as follows

$$\hat{\beta}^{\text{LASSO}}(\lambda) = \underset{\beta}{\text{argmin}}\left(\frac{\|Y-X\beta\|_2^2}{n} + \lambda\|\beta\|_1\right), \qquad (6)$$

where $\|\beta\|_1 = \sum_{j=1}^{k}|\beta_j|$ and $\lambda > 0$ is the parameter that controls the strength of the penalty. The larger value of λ, means the greater in the amount of shrinkage. The relation between λ and the upper bound t is a reverse relationship as t tends to ∞, the problem becomes an ordinary least square and λ equals 0. Vice versa as t equals 0, all coefficients shrink to 0 and λ tends to ∞. When minimizing the optimization problem some coefficients are shrank to zero, i.e. $\hat{\beta}(\lambda) = 0$,

for some values of j (depending on the value of the parameter λ). In this case the explanatory variables with coefficients equal to zero are excluded from the model.

2.3 Elastic Net Regression

The EN is a regularized regression method which overcomes the limitations of the LASSO. This method is very useful when $k > n$ or there are many correlated variables. The EN as a generalization of the LASSO, to be a valuable tool for model fitting and feature extraction.

The EN criterion is defined by

$$\min \|Y^* - X^*\beta\|_2^2$$

$$\text{subject to } \|\beta\|_2^2 \le t$$

$$\|B\|_1 \le t, \qquad (7)$$

$$\hat{\beta}_{EN}(\lambda_1, \lambda_2) = \min_{\beta}\Big(\sum_{i=1}^{n}(y_i - \sum_{j=1}^{k}\beta_j x_{ij})^2 + \lambda_1 \sum_{j=1}^{k}|\beta_j| + \lambda_2 \sum_{j=1}^{k}\beta_j^2\Big) \ . \ (8)$$

Which depends on two regularized parameters $\lambda_1, \lambda_2 > 0$. The EN penalty is a convex combination of the LASSO and ridge penalty and, in constraint form, it is given by

$(1-\alpha)\sum_{j=1}^{k}|\beta_j| + \alpha\sum_{j=1}^{k}\beta_j^2 \le t$ with $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$.

When $\alpha = 1$ the EN tend to RR and, at $\alpha = 0$ the EN tend to LASSO. There are advantages of the EN given as follows:

- A group of correlated variables can be selected without arbitrary omissions.
- The number of selected variables is no longer limited by the sample size [Karl and Simar, (2015)].

This paper is concerned with introducing regularization and selection of variables method as a

generalization of the LASSO which combines the two penalties λ_1 and λ_2 . This objective depend on the λ_1 realizes which is the advantages of the LASSO method and λ_2 realizes is tent to the advantages of the ridge method. The generalized method will be applied to Low-dimensional data and High-dimensional data and will be compare to the LASSO and ridge method.

3.The Monte Carlo Simulation Study

In this section, a Monte Carlo simulation study is conducted to evaluate and compare the performance of (OLS, RR, LASSO and, EN) estimators. Programs R are used to solve the numerical analysis by some packages (MASS, LM ridge, …

The simulated dataset is carried out based on eq.(1), with the following simulation settings:

1. Different sample sizes are selected as: $n = 75, 100, 200\ and, 300$ in the case of Low-dimensional data and $n = 25,50,100,150$ in the case of High-dimensional data.

2. Different number of the explanatory variables as $k = 6\ and, 20$ for Low-dimensional datasets, while $k = 100, 200\ and, 300$ for High-dimensional in datasets.

3. The true value of the intercept equals one ($\beta_0 = 1$), and the true values of β as:$(\beta_1, \dots, \beta_{k/2})' = 5$, and $\left(\beta_{\left(\frac{k}{2}\right)+1}, \dots, \beta_k\right)' = 0$.

When using models which depend on choosing explanatory variables the used method in the simulation aim to determine the significant variables in which half of the explanatory variables can be set as non-significant to test the mechanism for selecting the explanatory variables in the proposed model

used in the present simulation and the non-significant ones are discarded from the simulation.

4. The explanatory variables are generated from multivariate normal $MVN\,(1,\Sigma_x)$, where diag (Σ_x) = 1 and off-diag (Σ_x) = ρ_x, where $\rho_x = 0$ and 0.95. Note that when $\rho_x = 0.95$this means that the model has a multicollinearity problem.

5. The error is generated from the standard MND.

6. For generating some outlier values in the model, some values are replaced randomly (according to the selected ratio of outlier values: 0% and 15%) in y with other values generated from MND as $N(mean(y) + 5 * IQR(y),1)$ where IQR is inter quartile range.

7. All Monte Carlo experiments involved 500 replications and all the results of all separate experiments are obtained by precisely the same series of random numbers.

In this paper, the root mean squared error (RMSE), mean absolute error (MAE) and $R^2$are used as the criteria of judgment:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{n - K}} \quad , \quad \mathrm{MAE} = \frac{1}{n - K}\sum_{i=1}^{n}|\hat{y}_i - \bar{y}_i|,$$

$$R^2 = \frac{\hat{\beta}\dot{X}\dot{y} - n\bar{y}^2}{\dot{y}\dot{y} - n\bar{y}^2}.$$

The simulation results are recorded in Tables 1 – 22. These Tables represent the RMSE, MAE and,R^2 when the selected variables for the estimators in the different factors (k, ρ_x, and outliers). From tables 1 – 16 represent the results of the various estimators in case of Low-dimensional data. From tables 17 – 22 represent the results of the various estimators

in case of High-dimensional data. The LASSO and EN methods are used in High- dimensional.

Table (1): The average estimators of RMSE, MAE, and, R^2 for OLS, RR, LASSO and, EN **when there are no outliers = 0%, n = 75 and k=6**

ρ_{ε}	Method	RMSE	MAE	R^2	Selected variables
0.00	OLS	1.08774	0.87559	0.98912	ALL
	RR	1.28059	1.03223	0.98882	ALL
	LASSO	1.07173	0.86166	0.98959	5
	EN	1.07347	0.86338	0.98960	5
0.95	OLS	1.09073	0.87878	0.99529	ALL
	RR	1.47048	1.18280	0.99175	ALL
	LASSO	1.06753	0.86040	0.99553	4
	EN	1.06228	0.85582	0.99557	4

Table (2): The average estimators of RMSE, MAE, and, R^2 for OLS, RR, LASSO and, EN **when outliers = 0%, n = 100 and k = 6**

ρ_{ε}	Method	RMSE	MAE	R^2	Selected variables
0.00	OLS	1.06406	0.85516	0.98911	ALL
	RR	1.24228	0.99892	0.98891	ALL
	LASSO	1.04973	0.84333	0.98951	5
	EN	1.05521	0.84755	0.98938	5
0.95	OLS	1.07274	0.86261	0.99578	ALL
	RR	1.46977	1.17778	0.99238	ALL
	LASSO	1.06127	0.85477	0.99594	4
	EN	1.05887	0.85224	0.99594	4

Table (3): The average estimators of RMSE, MAE, and, R^2 for OLS, RR, LASSO and, EN when outliers = 0%, n = 200 and k = 6

ρ_ε	Method	RMSE	MAE	R^2	Selected variables
0.00	OLS	1.03000	0.82500	0.98884	ALL
	RR	1.18740	0.95025	0.98872	ALL
	LASSO	1.02429	0.82036	0.98901	5
	EN	1.02478	0.82083	0.98901	5
0.95	OLS	1.01405	0.81164	0.99520	ALL
	RR	1.35902	1.09159	0.99159	ALL
	LASSO	1.00775	0.80639	0.99525	4
	EN	1.01216	0.81108	0.99526	4

Table (4): The average estimators of RMSE, MAE, and, R^2 for OLS, RR, LASSO and, EN when outliers = 0%, n = 300 and k = 6

ρ_ε	Method	RMSE	MAE	R^2	Selected variables
0.00	OLS	1.02412	0.81885	0.98807	ALL
	RR	1.17060	0.93508	0.98796	ALL
	LASSO	1.02029	0.81575	0.98817	5
	EN	1.02077	0.81626	0.98817	5
0.95	OLS	1.02320	0.81795	0.99552	ALL
	RR	1.40080	1.11836	0.99193	ALL
	LASSO	1.02041	0.81579	0.99560	4
	EN	1.01983	0.81522	0.99558	4

Tables 1 – 4 represent the result of the obtained estimators from applying the methods of OLS, ridge , LASSO and EN when there are no outliers using different

sample sizes as n= 75, 100,200,300 and number of explanatory variables is k = 6 in the case of low- dimensional data. When the coefficient of correlation ρ_x is equal to zero (there is no collinearity between the explanatory variables) the LASSO method has the smallest values for the measures RMSE, MAE. But when the value of $\rho_x = 0.95$ (there is collinearity between the explanatory variables) it is noticed that the EN method has the smallest values for the measures RMSE, MAE. Also, the EN has the greatest value for R^2 with the two values of ρ_x which means that this method can explains more variation which affect the dependent variable. In addition ,LASSO and EN are appropriate in selecting variables.

Table (5): The average estimators of RMSE, MAE, and, R^2 for OLS, RR, LASSO and, EN when outliers = 0%, n = 75 and k = 20

ρ_x	Method	RMSE	MAE	R^2	Selected variables
0.00	OLS	1.36559	1.10059	0.99156	ALL
	RR	2.18326	1.75484	0.98539	ALL
	LASSO	1.29168	1.03990	0.99269	17
	EN	1.30642	1.05249	0.99243	18
0.95	OLS	1.35964	1.09464	0.99930	ALL
	RR	2.18177	1.75228	0.99819	ALL
	LASSO	1.90169	1.52565	0.99938	13
	EN	1.83949	1.47497	0.99935	13

Table (6): The average estimators of RMSE, MAE, and, R^2 for OLS, RR, LASSO and, EN when outliers = 0%, n = 100 and k = 20

ρ_{ε}	Method	RMSE	MAE	R^2	Selected variables
0.00	OLS	1.24789	1.00096	0.99280	ALL
	RR	1.82679	1.46354	0.99023	ALL
	LASSO	1.20555	0.96734	0.99344	17
	EN	1.21485	0.97476	0.99326	18
0.95	OLS	1.24427	1.00038	0.99935	ALL
	RR	2.00654	1.60536	0.99832	ALL
	LASSO	1.73659	1.39706	0.99939	13
	EN	1.69774	1.36547	0.99937	14

Table (7): The average estimators of RMSE, MAE, and, R^2 for OLS, RR, LASSO and, EN when outliers = 0%, n = 200 and k = 20

ρ_{ε}	Method	RMSE	MAE	R^2	Selected variables
0.00	OLS	1.10977	0.88736	0.99514	ALL
	RR	1.37944	1.10730	0.99479	ALL
	LASSO	1.08996	0.87211	0.99537	16
	EN	1.09219	0.87356	0.99532	17
0.95	OLS	1.10351	0.88335	0.99949	ALL
	RR	2.09226	1.67573	0.99821	ALL
	LASSO	1.62510	1.29316	0.99953	12
	EN	1.58834	1.26406	0.99952	12

Table (8): The average estimators of RMSE, MAE, and, R^2 for OLS, RR, LASSO and, EN when outliers = 0%, n = 300 and k = 20

ρ_x	Method	RMSE	MAE	R^2	Selected variables
0.00	OLS	1.07169	0.85724	0.99508	ALL
	RR	1.31049	1.05029	0.99487	ALL
	LASSO	1.05941	0.84737	0.99524	16
	EN	1.06123	0.84877	0.99521	17
0.95	OLS	1.06710	0.85376	0.99953	ALL
	RR	2.03264	1.61117	0.99830	ALL
	LASSO	1.59312	1.26838	0.99955	12
	EN	1.56076	1.24245	0.99954	12

Tables 5 - 8 represent the results of the obtained estimators from applying the methods of OLS, ridge, LASSO and EN when there are no outliers using different sample sizes as $n = 75, 100, 200, 300$ and the number of explanatory variables is $k = 20$ in the case of low- dimensional data. when the coefficient of correlation ρ_x is equal to zero (there is no collinearity between the explanatory variables) the LASSO method has the smallest values for the measures RMSE, MAE. But, when the value of $\rho_x = 0.95$ (there is collinearity between the explanatory variables) the EN method has the smallest values for the measures RMSE, MAE. In addition, the LASSO has the compatible value of R^2, also LASSO and EN is sufficient in selecting variables.

Table(9): The average estimators of RMSE, MAE, and, R^2 for OLS, RR, LASSO and, EN when outliers = 15%, n = 75,and k= 6

ρ_ε	Method	RMSE	MAE	R^2	Selected variables
0.00	OLS	21.96185	15.54582	0.14867	ALL
	RR	21.79081	15.35956	0.14933	ALL
	LASSO	21.25412	14.99399	0.15243	3
	EN	21.33063	14.97304	0.15203	5
0.95	OLS	42.91193	30.51059	0.08701	ALL
	RR	40.89547	28.47921	0.11096	ALL
	LASSO	40.20925	28.15167	0.13374	2
	EN	39.92546	28.00635	0.14295	3

Table (10): The average estimators of RMSE, MAE, and, R^2 for OLS, RR, LASSO and, EN when outliers = 15%, n = 100 and k= 6

ρ_ε	Method	RMSE	MAE	R^2	Selected variables
0.00	OLS	22.89620	16.14534	0.11231	ALL
	RR	22.78355	16.04157	0.11350	ALL
	LASSO	22.42343	15.81424	0.11605	4
	EN	22.41558	15.76245	0.11604	5
0.95	OLS	34.71381	24.50302	0.10930	ALL
	RR	33.39004	23.41353	0.13821	ALL
	LASSO	33.10362	23.35253	0.15799	2
	EN	32.96437	23.22943	0.16496	3

Table (11): The average estimators of RMSE, MAE, and, R^2 for OLS, RR, LASSO and, EN when outliers = 15%, n = 200 and k= 6

$\rho_\perp$	Method	RMSE	MAE	R^2	Selected variables
0.00	OLS	22.42146	15.64623	0.10198	ALL
	RR	22.36787	15.60029	0.10197	ALL
	LASSO	22.26792	15.58305	0.10407	4
	EN	22.22338	15.55020	0.10409	5
0.95	OLS	39.45112	27.65671	0.07698	ALL
	RR	38.95378	27.38085	0.08821	ALL
	LASSO	38.77864	27.31436	0.09990	2
	EN	38.64149	27.23241	0.10601	3

Table(12): The average estimators of RMSE, MAE, and, R^2 for OLS, RR, LASSO and, EN when outliers = 15%, n = 300 and k= 6

$\rho_\perp$	Method	RMSE	MAE	R^2	Selected variables
0.00	OLS	22.19938	15.48367	0.10129	ALL
	RR	22.16674	15.46751	0.10179	ALL
	LASSO	22.12442	15.47802	0.10288	5
	EN	22.10581	15.47234	0.10267	5
0.95	OLS	35.87823	25.05733	0.09712	ALL
	RR	35.49692	24.92147	0.10984	ALL
	LASSO	35.36313	24.87376	0.11689	3
	EN	35.35131	24.89073	0.11902	4

Tables 9 - 12 represent the results of the obtained estimators from applying the methods of OLS, ridge, LASSO and EN when there are 15% outliers using different sample sizes as n =75, 100, 200, 300 and number of explanatory variables is k = 6 in the case of low-dimensional data. when the coefficient of correlation $\rho_x$is equal to zero (there is no collinearity between the explanatory variables) the EN method has the smallest values for the measures RMSE, MAE. But, when the value of $\rho_x = 0.95$ (there is collinearity between the explanatory variables) the EN method has the smallest values for the measures RMSE, MAE. In addition, EN has the compatible value of R^2, also LASSO and EN are appropriate in selecting variables.

Table (13): The average estimators of RMSE, MAE, and, R^2 for OLS, RR, LASSO and, EN when outliers = 15%, n = 75and k= 20

ρ_x	Method	RMSE	MAE	R^2	Selected variables
0.00	OLS	50.56980	38.80567	0.05190	ALL
	RR	48.36526	36.67422	0.05376	ALL
	LASSO	39.46258	28.48060	0.05380	3
	EN	42.44671	31.09004	0.05325	13
0.95	OLS	148.62944	114.21679	0.05886	ALL
	RR	118.25861	83.64788	0.09943	ALL
	LASSO	110.84558	77.68443	0.15016	3
	EN	110.20993	77.07346	0.15896	4

Table (14): The average estimators of RMSE, MAE, and, R^2 for OLS, RR, LASSO and, EN when outliers = 15%, n = 100 and k= 20

ρ_ε	Method	RMSE	MAE	R^2	Selected variables
0.00	OLS	46.23871	34.74597	0.04655	ALL
	RR	45.12440	33.68611	0.04570	ALL
	LASSO	39.66431	28.58847	0.04235	5
	EN	41.58705	30.19934	0.04496	13
0.95	OLS	138.93487	104.65616	0.05310	ALL
	RR	120.41259	85.59050	0.08220	ALL
	LASSO	113.91253	80.30376	0.12335	3
	EN	113.33949	79.88300	0.13192	4

Table (15): The average estimators of RMSE, MAE, and, R^2 for OLS, RR, LASSO and, EN when outliers = 15%, n = 200 and k= 20

ρ_ε	Method	RMSE	MAE	R^2	Selected variables
0.00	OLS	43.12061	30.97939	0.05354	ALL
	RR	42.79971	30.67705	0.05372	ALL
	LASSO	41.05434	29.10741	0.05071	10
	EN	41.36831	29.27857	0.05111	14
0.95	OLS	116.60006	83.67337	0.06392	ALL
	RR	109.27017	76.23678	0.09564	ALL
	LASSO	106.66929	74.76825	0.12690	4
	EN	105.90301	74.52763	0.13708	5

Table (16): The average estimators of RMSE, MAE, and, R^2 for OLS, RR, LASSO and, EN when outliers = 15%, n = 300 and k= 20

ρ_x	Method	RMSE	MAE	R^2	Selected variables
0.00	OLS	40.88605	28.84036	0.05755	ALL
	RR	40.71706	28.71606	0.05713	ALL
	LASSO	39.85579	28.04578	0.05414	12
	EN	39.91478	28.04950	0.05344	14
0.95	OLS	139.23361	98.33311	0.04825	ALL
	RR	133.52371	93.27302	0.07049	ALL
	LASSO	131.14743	92.26240	0.09307	4
	EN	130.65465	92.19347	0.10038	5

Tables 13 - 16 represent the results of the obtained estimators from applying the methods of OLS, ridge, LASSO and EN when there are 15% of outliers using different sample sizes as n = 75, 100, 200, 300 and the number of explanatory variables is k = 20 in the case of low-dimensional data. when the coefficient of correlation ρ_x is equal to zero (there is no collinearity between the explanatory variables) the LASSO method has the smallest values for the measures RMSE, MAE. But, when the value of $\rho_x = 0.95$ (there is collinearity between the explanatory variables) the EN method has the smallest values for the measures RMSE, MAE. In addition, EN has the compatible value of R^2 and it are appropriate in selecting variables.

Table(17): The average estimators of RMSE, MAE, and, R^2 for LASSO and EN for HD data when outliers = 0% and k= 100

n	ρ_{ϵ}	Method	RMSE	MAE	R^2	Selected variables
25	0.00	LASSO	25.06338	21.42706	0.09941	1
		EN	26.06660	22.14841	0.09422	35
	0.95	LASSO	23.25811	19.20100	0.99576	29
		EN	10.46502	8.47863	0.99959	100
50	0.00	LASSO	33.87210	26.97234	0.14755	28
		EN	29.91881	23.96333	0.27175	62
	0.95	LASSO	18.12829	15.03965	0.99701	43
		EN	9.01871	7.48916	0.99971	100

Table (18): The average estimators of RMSE, MAE, and, R^2 for LASSO and EN for HD data when outliers = 0% and k= 200

n	ρ_{ϵ}	Method	RMSE	MAE	R^2	Selected variables
50	0.00	LASSO	55.35134	43.85659	0.04847	1
		EN	54.41400	43.57338	0.07972	83
	0.95	LASSO	36.22438	29.63382	0.99431	54
		EN	15.17606	12.40342	0.99960	200
100	0.00	LASSO	47.49083	38.94290	0.15968	43
		EN	45.05417	36.89456	0.20699	99
	0.95	LASSO	29.13738	23.76127	0.99832	92
		EN	16.99362	14.08550	0.99982	200

Table(19): The average estimators of RMSE, MAE, and, R^2 for LASSO and EN for HD data when outliers = 0% and k= 300

n	ρ_x	Method	RMSE	MAE	R^2	Selected variables
100	0.00	LASSO	58.78577	47.58312	0.12074	44
		EN	55.90233	45.48068	0.17546	143
	0.95	LASSO	44.77200	36.44241	0.99822	101
		EN	25.44278	21.03140	0.99988	300
150	0.00	LASSO	60.00250	47.54373	0.13709	144
		EN	53.36222	42.09911	0.21074	168
	0.95	LASSO	37.54366	30.54349	0.99857	124
		EN	23.36873	18.42799	0.99984	300

Tables 17 - 19 represent the result of the obtained estimators from applying the methods of LASSO and EN when there are no outliers in the data using different sample sizes, the coefficient of correlation $(\rho_x = 0 \text{ and } 0.95)$ and different number of explanatory variables in the case of high-dimensional data. When the sample size is increased and when $(\rho_x = 0 \text{ and } 0.95)$ between the explanatory variables it is noticed that EN method has the smallest values for the measures of RMSE and MAE which means it is the best method than LASSO method. Also, EN method has the greatest value of R2 than LASSO method which means that explains more variations than LASSO method and it is better in selecting variables.

Table (20): The average estimators of RMSE, MAE, and, R^2 for LASSO and EN for HD data when outliers = 15% and k= 100

n	ρ_ε	Method	RMSE	MAE	R^2	Selected variables
25	0.00	LASSO	72.21272	51.62750	0.13598	4
		EN	72.27706	51.97192	0.14042	22
	0.95	LASSO	484.99485	334.39289	0.35742	4
		EN	472.99445	323.91423	0.38549	17
50	0.00	LASSO	106.27043	78.39640	0.06370	6
		EN	106.60289	78.61635	0.06826	14
	0.95	LASSO	655.28613	455.06672	0.15213	5
		EN	641.76743	444.11853	0.17677	16

Table(21): The average estimators of RMSE, MAE, and, R^2 for LASSO and EN for HD data when outliers = 15% and k= 200

n	ρ_ε	Method	RMSE	MAE	R^2	Selected variables
50	0.00	LASSO	124.47083	92.72937	0.06354	7
		EN	125.08160	93.31391	0.06824	25
	0.95	LASSO	1076.60643	748.09731	0.17620	7
		EN	1053.37536	729.24622	0.20620	28
100	0.00	LASSO	140.24401	101.35687	0.03012	6
		EN	142.60531	103.46130	0.03146	13
	0.95	LASSO	1209.49637	849.51524	0.12216	7
		EN	1191.96701	838.94223	0.14405	21

Table (22): The average estimators of RMSE, MAE, and, R^2 for LASSO and EN for HD data when outliers = 15% and k= 300

n	ρ_x	Method	RMSE	MAE	R^2	Selected variables
100	0.00	LASSO	174.31765	126.80075	0.03104	7
		EN	176.63074	128.99068	0.03233	16
	0.95	LASSO	1914.42992	1346.29735	0.11670	8
		EN	1886.09733	1326.86802	0.13884	28
150	0.00	LASSO	151.88982	110.03227	0.01951	8
		EN	153.79148	111.72149	0.01946	15
	0.95	LASSO	1794.79567	1247.58636	0.10453	7
		EN	1771.07779	1235.90145	0.12219	19

Tables 20 - 22 represent the result of the obtained estimators from applying the methods of LASSO and EN when there are 15% outliers using different sample sizes, the coefficient of correlation $(\rho_x = 0 \text{ and } 0.95)$ and different number of explanatory variables in the case of high-dimensional data. When data has outliers, $(\rho_x = 0)$ and the sample size is increased reflects the RMSE, MAE and R2 value of LASSO and EN methods are relatively close, and the EN method is better in selecting variables. But, when the value of $\rho_x = 0.95$ the EN method has the smallest values for the measures RMSE, MAE. In addition, EN has the greatest value of R^2 and it is better in selecting variables.

4.Conclusions

The simulation results indicate that the obtained estimators using EN are efficient and reliable than the other

estimators when the explanatory variables are highly correlated as well as the dataset contains outliers or not. It is concluded that the values of RMSE and MAE of EN estimators are smaller than the RMSE and MAE for the other estimators in both the two cases of low and high dimensional data. In addition, the EN method is better than LASSO when it used to be selecting variables.

In this paper, the performance of EN estimator compared to and OLS, ridge, and LASSO estimators under different situations. The evaluation is based on using the measures of RMSE, MAE, andR^2. Two cases are considered: low-dimensional data and high-dimensional data based on providing an introduction of the regression analysis, illustration the penalized regression methods of most common. The Monte Carlo simulation study is conducted to evaluate and compare the performance of these estimators under different situations. The simulation results indicate that the obtained estimators using EN are efficient and reliable than the other estimators when the explanatory variables are highly correlated. Finally, we can say that the LASSO method helps us to choose a model with the most relevant features in general, it is advised to use the EN method with multicollinearity problems and high-dimensional datasets.

References

1. Adelegoke, A.Adewuyi, E.Ayinde,K. and Lukman,A. (2016). "A Comparative Study of some Robust Ridge and Liu Estimators", *Science World journal* , Vol. 11, pp.16-20.
2. Alauddin,M. and Nghiemb,H. (2010). "Do Instructional Attributes Pose Multicollinearity Problems? An Empirical Exploration." *Economic Analysis and Policy*, Vol. 40, pp. 351–361.
3. Alkan, B.B, Atakan, C. (2013). Visualizing Diagnostic of Multicollinearity: Table Plot and Biplot Methods. *Pakistan Journal of statistics*, Vol. 29, pp. 59-78.
4. Arslan, O. (2012). Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression. Computational Statistics and Data Analysis, Vol 56, 1952– 1965.
5. Emmert-Streib, F. and Dehmer,M. (2019). High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection. *Machine Learning and knowledge Extraction*, Vol. 1, pp. 359-383.
6. Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. JASA,Vol 96, 1348-1360.
7. Fonti,V. and Belitser, E. (2017). "Feature Selection Using LASSO", *Vrije Universiteit Amsterdam*.
8. Hebiri,M. and Van,S. (2011). "The Smooth-Lasso and other $l_1 + l_2$-Penalized Methods". arXiv:1003.4885v2 [math.ST]
9. Hoerl, A. E. (1962). Application of Ridge Analysis to Regression Problems. *Chemical Engineering Progress*, Vol 58, no 3, pp. 54-59.
10. Hoerl, A.E. and Kennard, R. W. (1970a). Ridge Regression: Biased Estimation for Non orthogonal Problems. *Technometrics*, Vol. 12, no. 1, pp. 55 - 67.
11. Hoerl, A.E. and Kennard, R. W. (1970b). Ridge Regression: Application to Non-Orthogonal Problems. *Technometrics*, vol. 12, no. 3, pp. 591 - 612.

12. Huang, J., Jiao, Y., Liu, Y. and Lu, X. in (2018). A Constructive Approach to l_0Penalized Regression, *Machine Learning Research*, *Tong Zhang.*
13. Hui and Zou. (2006). "The Adaptive Lasso and its Oracle Properties". *Journal of the American statistical association*, Vol. 101, no. 476, pp.1418–1429.
14. Januaviani,T. M., Gusriani, N., Subiyanto and Bon,A. T. (2019). "The LASSO (Least Absolute Shrinkage and Selection Operator) Method to Predict Indonesian Foreign Exchange Deposit Data*" Proceedings of the International Conference on Industrial Engineering and Operations Management* Bangkok, Thailand. pp.3195-3202
15. James,G.Witten,D. Hastie,T. and Tibshirani,R. (2013). An Introduction to Statistical Learning with Applications in R. Springer New York.
16. Karl.W and Simar.L (2015). Applied Multivariate Statistical Analysis, Springer.
17. Melkumova, L. E. and Shatskikh, S.Y. (2017). "Comparing Ridge and LASSO Estimators for Data Analysis", *Procedia Engineering* 201 pp.746–755
18. Neter, J., Wasserman,W. and Kutner,M.H. (2005). Applied Linear Statistical Models. McGraw-Hill/Irwin New York.
19. Osborne M, Presnell B, Turlach B (2000). "A New Approach to Variable Selection in LeastSquares Problems." *IMA Journal of Numerical Analysis*, Vol 20, p.p 389–404.
20. Saleh,E., Arashi,M. and Kibria,G. (2019). Theory of Ridge Regression Estimation with Applications. WILEY.
21. Tibshirani, R. (1996). "Regression Shrinkage and Selection Via the Lasso". *Journal of the Royal Statistics*. Vol. 58, pp. 267-288.
22. Tibshirani, R. (1997). The Lasso Method for Variable Selection in The COX Model. *Statistics in Medicine*, Vol. 16, pp.385—395.
23. Weige, Yu.*et al.* (2019). "Analysis of Vegetable Price Fluctuation Law and Causes based on Lasso Regression Model". Journal *of Physics: Conference Series*.

24. Zou,H., Hastie,T. (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol, Vol 67, p.p 301–320.

25. Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association, Vol 101, p.p 1418–1429.