The Effect of Outlier on Lasso Estimators and Regressions

Layla M. Nassir[*1]
Assistant Professor Dr., Mechanical Engineering Dept. College of Eng., Al-Mustansiriyah University, Baghdad, Iraq
Email: layla_matter@uomustansiriyah.edu.iq

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract:
Lasso regression (Least Absolute Shrinkage and Selection Operator) dependent on reducing shrinkage. This kind regression deals with cases in which the explained variables have a multicollinearity problem between them and in models include a large number of explained variables with the goal is to focus on the variables that have the most effect on the dependent variable. In this research Lasso regression were presented with deferent (sample size, number of explained variables and number of outliers) to show its effect on lasso and Bayesian lasso regression. Numerical results showed that Lasso estimator was affected by each of the sample size, outlier's ratios and regression method. Other methods, such as shrinkage ridge and Bayesian ridge methods can be used for comparison with the assumed methods.

Key words: Lasso Regression, Bayesian Lasso Regression, Explained Variables, Mean Square Error, Multicollinearity Problem, Outliers.

1.Introduction

Lasso Regression It is a new variable selection technique proposed by Robert Tibshirani in 1996. Lasso is a method for reducing shrinkage. Its basic idea is to reduce the sum of the squared residuals under the constraint that the sum of the absolute error values of the regression coefficients is less than a constant

In this field, many researches have been proposed like:-

The research presented by Jian Huang and others in (2006) In this paper, the effect of regression has been shown in the case of many dimensions of the explained variables With the change of covariance when increasing the sample size[4].

The research presented by Yiyuan she and others in (2011). The research include Nonconvex Penalized to detect outlier values in regression data sets with Masking in various P values [8]

The research presented by S.M.A.Khaleelur Rahman and others in (2012). The research include the Outliers are affected by multiple regression according to the least squares method The results show that the outliers method was affected [5].

The research presented by Achim Ahrens and others in (2019). The research includes the method of least squares and its effect on the number of explained variables, while presenting the theoretical aspects of lasso and ridge regression [1].

This research include applying Lasso regression on deferent data sets with outlier ratios.

2.Lasso Reg

Lasso method was first presented in geophysical literature in (1982) [8], The term Lasso represents the first letters of the concept (Least Absolute Shrinkage and Selection Operator), it is a penalty function of the a method for estimating the parameters of the regression model and selecting with organizing the variables included in the model to increase the explanatory accuracy of the regression models by choose a subset of the common variables in the final model instead of using all of them, in the Lasso method the sum square errors of the proposed model is minimized [3].

Lasso was originally designed for Least squares models with a large amount of estimator behavior via the Lasso parameter or so-called Soft Thresholding, including the relationship of the Lasso estimator with the Ridge Regression estimator and the best subset selection of the variables. Which is similar to the Stepwise selection method, Lasso coefficient estimates should not be single if the explanatory variables suffer from the problem of multicollinearity

Lasso method has the ability to choose a subset based on the constraint formula, and although the Lasso is defined for least squares, the Lasso method can easily be used in a wide range of statistical models, including generalized linear models, generalized estimation factors, relative risk models, and

M estimators. Lasso can be used in many fields such as geometry, Bayesian statistics, and convex analysis

Before the Lasso regression method, the most used method for selecting the explanatory variables that are included within the model was the Stepwise Selection method, which improves the accuracy of the model in certain cases, especially when some explanatory variables have a strong relationship with the response variable, which makes the prediction inaccurate, as well as Ridge Regression is the most popular method used to improve the prediction accuracy of the regression model. It improves prediction error by reducing large regression coefficients in order to reduce redundancy, but does not perform co-selection and thus does not help make the model more interpretable.

Whereas Lasso can achieve both goals by making the set of absolute values of the regression coefficients have quantities less than a constant value, forcing some of the coefficients to be equal to zero, while choosing a simpler model that does not include these coefficients.

1-2 General Lasso Formula[6,7,9]

Lasso regression parameters were estimated according to the principle of least squares from the basic formula as follows:-

$$\min\left\{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(y_i-\upbeta_0-\mathrm{x_i}^{\mathrm{T}}\upbeta)^2\right\}\ \dots(1)$$

Subject to $\sum_{\mathrm{j}=1}^{\mathrm{p}}|\upbeta_{\mathrm{j}}|\leq \mathrm{t}$

With

$N\ \ reresent\ sampel\ size$

$Y\ \ represent\ dependent\ variabel\ with\ size\ NX1$

$X\ \ represent\ Explained\ Variables\ with\ size\ NXP$

$P\ \ represent\ number\ of\ Explained\ Variables$

$t\ \ represent\ \ a\ pre-set\ free\ parameter\ that\ specifies\ the\ amount\ of\ shrink$

Lasso formula can be written as follows:-

$$\min_{\upbeta_0,\upbeta}\{\frac{1}{\mathrm{N}}||\mathrm{y}-\mathrm{X}\upbeta||_2^2\}\quad \dots(2)$$

Subject to $||\upbeta||\leq \mathrm{t}$

With

$$||\upbeta||_{\mathrm{p}}=(\sum_{\mathrm{i}=1}^{\mathrm{N}}|\upbeta_{\mathrm{i}}|^{\mathrm{p}})^{1/\mathrm{p}}\qquad \dots(3)$$

When (P=1) Then ($||\upbeta||_1$) becomes the standard length (ℓ^{p})

Since we have

$$\hat{\upbeta}_0=\bar{\mathrm{y}}-\bar{\mathrm{x}}_{\mathrm{i}}^{\mathrm{T}}\,\upbeta\ \ \dots(4)$$

Then

$$\mathrm{yi}-\upbeta_0-\mathrm{x_i}^{\mathrm{T}}\upbeta=\mathrm{yi}-\left(\bar{\mathrm{y}}-\bar{\mathrm{x}}_{\mathrm{i}}^{\mathrm{T}}\,\upbeta\right)-\mathrm{x_i}^{\mathrm{T}}\upbeta$$
$$\mathrm{yi}-\upbeta_0-\mathrm{x_i}^{\mathrm{T}}\upbeta=(\mathrm{yi}-\bar{\mathrm{y}})-(\mathrm{x_i}-\bar{\mathrm{x}})^{\mathrm{T}}\upbeta$$

with

($\bar{\mathrm{x}}$) denotes the standard mean of the data points ($\mathrm{x_i}$)

($\bar{\mathrm{y}}$) the mean of the dependent variable (response variable (yi))

Thus, it is natural to work with variables that have been centralized (making their mean equal to zero) in addition to the explanatory variables being Typically standardizes

i.e.

$$\frac{1}{N}\sum_{\mathrm{i}=1}^{\mathrm{N}}\mathrm{x_i}=0\ \text{ and }\ \frac{1}{N}\sum_{\mathrm{i}=1}^{\mathrm{N}}\mathrm{x_i}^2=1$$

Then formula (1) can be rewritten as follows:-

$$\min_{\upbeta_0,\upbeta}\{\frac{1}{\mathrm{N}}||\mathrm{y}-\mathrm{X}\upbeta||_2^2\}\quad \dots(5)$$

Subject to

$||\upbeta||_1\leq \mathrm{t}$

It is in the LaGrange multiplicative form of the as follows:-

$$\min_{\beta\in R^P}\{\frac{1}{N}||y - X\beta||_2^2 + \lambda||\beta||_1\}$$

With

(λ) denote the parameter that controls the penalty force (shrinkage) over the regression estimators.

2.2 Properties of Lasso Estimators[4]

There are some lasso estimator properties that can list as follows:-

a- Orthonormal Covariates

Suppose that covariates are normally orthogonal, such that

$$\left(x_i \middle| x_j\right) = \delta_{ij}$$

With

$(.|.)$ *denote Inner product*

δ_{ij} *denote Kroncher delta* Such that

$$\delta_{ij} = \begin{cases} 0 & if \ \ i \neq j \\ 1 & if \ \ i = j \end{cases}$$

By using the iterative method of Sub gradient, which is one of the methods for solving less intrusive problems, we obtain:-

$$\hat{\beta}_j = S_{N\lambda}\left(\hat{\beta}_j^{\ OLS}\right) = \hat{\beta}_j^{\ OLS}\text{Max}(0,1 - \frac{N\lambda}{\left|\hat{\beta}_j^{\ OLS}\right|}) \quad \dots (6)$$

With

$$\hat{\beta}_j^{\ OLS} = (X^T X)^{-1} X^T Y \quad \dots (7)$$

and

$S_{N\lambda}$ *denote Smooth Threshold*

So the goal is to reduce the following formula

$$\min_{\beta\in R^P}\{\frac{1}{N}\left|\left|y - X\beta\right|\right|_2^2 + \lambda\left|\left|\beta\right|\right|_2^2\} \quad \dots (8)$$

And

$$\hat{\beta}_j = (1 + N\lambda)^{-1}\hat{\beta}_j^{\ OLS} \quad \dots (9)$$

since the ridge regression shrinks all the coefficients by the variable factor of

$$(1 + N\lambda)^{-1}$$

and does not put any of the coefficients to zero.

Then

$$\min_{\beta\in R^P}\{\frac{1}{N}\left|\left|y - X\beta\right|\right|_2^2 + \lambda\left|\left|\beta\right|\right|_0\} \quad \dots (10)$$

And

$$\hat{\beta}_j = H_{\sqrt{N\lambda}}\left(\hat{\beta}_j^{\ OLS}\right) = \left(\left|\hat{\beta}_j^{\ OLS}\right| > \sqrt{N\lambda}\right) \quad \dots (11)$$

With

$H_{\sqrt{N\lambda}}$ *represent Limit threshold*

We note that the Lasso estimates combine the characteristics of the ridge regression and the regression of the best partial choice. It converts all parameters to zero with a fixed value and adjusts them to zero if they reach.

b- Correlated Covariates

Returning to the general form of lasso in which the different covariates may not be explanatory, in which case two explanatory variables (i *and* j) are identical for each case so that

$x_i = x_j$

Then the parameter values

B_i & B_j

Which minimizes the Lasso objective function is not uniquely defined and in case of

$\hat{B}_i, \hat{B}_j \geq 0$

And if

$s \in [0,1]$

By replacing $(\hat{B}_i)$ by

$$s\left(\hat{B}_i + \hat{B}_j\right) \quad \& \quad 1 - s\left(\hat{B}_i + \hat{B}_j\right) \quad \dots (12)$$

While retaining the rest $(\hat{B}_i)$
And we'll get a new solution and the Lasso function continues to shrink the coefficients.

3. Bayesian Lasso Regression

Just like Bayesian approach the lasso regression estimated parameters have been assigned prior distribution to include in lasso regression, In (1996) Tibshirani use the lasso estimators as a posterior when the regression estimators distributed as independent identical Laplace distribution[2].
lasso estimators can proposed to be the mode of the posterior distribution for

$$\hat{\beta}_L = \arg max_\beta \; p\left(\frac{\beta}{y}, \sigma^2 \tau\right) \dots (13)$$

With

$$p\left(\frac{\beta}{\tau}\right) = \left(\frac{\tau}{2}\right)^p e^{-\tau\|\beta\|_1} \dots (14)$$

The likelihood

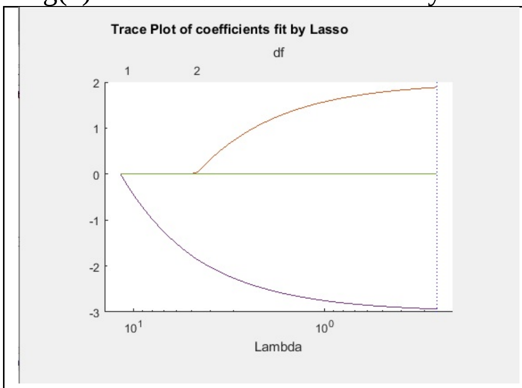
$$p\left(\frac{y}{\beta}, \sigma^2\right) = N\left(\frac{y}{XB}, \sigma^2 I_n\right) \dots (15)$$

For any fixed $(\sigma^2 > 0, \tau > 0)$ values the posterior mode for (β) will be lasso estimators with penalty function $(\lambda = 2\tau\sigma^2)$
The Bayesian Lasso is

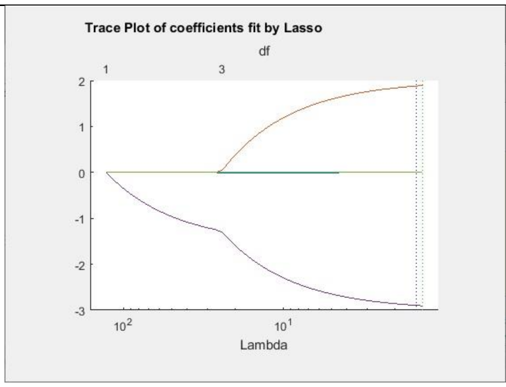
$$\pi\left(\frac{\beta}{\sigma^2}\right) = \frac{\lambda}{2\sigma} e^{\frac{-\frac{\lambda}{\beta_j}}{\sigma}} \qquad \dots (16)$$

4.Experimental Results

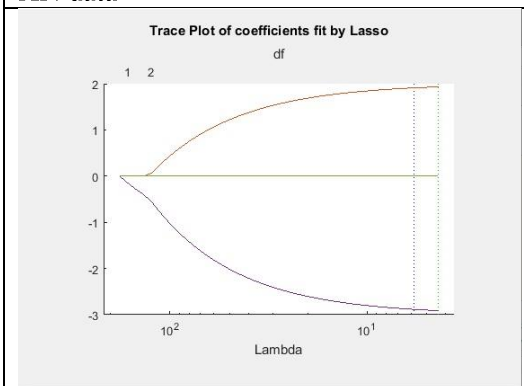
research included sets of data each of them with (3 Explained variables and 2000 sample size
$(XN \; with \; no \; Outlier \; , XN1 \; with \; 5\% \; Outliers, XN2 \; with \; 10\% \; Outliers,$
$XN3 \; with \; 15\% \; Outliers, XN4 \; with \; 20\% \; Outliers, XN5 \; with \; 25\% \; Outliers$)
The results are illustrated in the following figures and tables
Fig(1) the Plot of Coefficients fit by lasso with various Outlier ratios



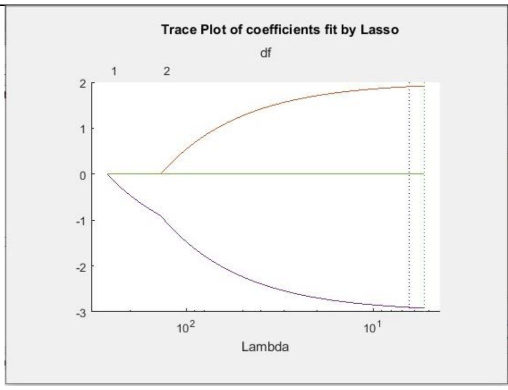
XN data



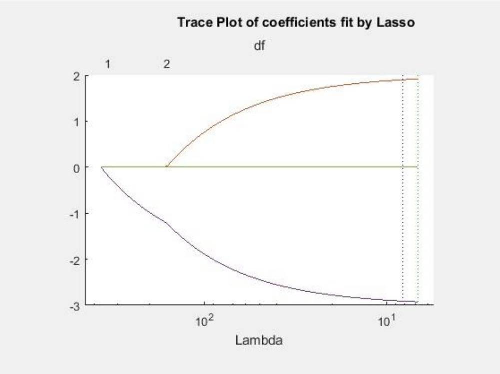
XN1 data



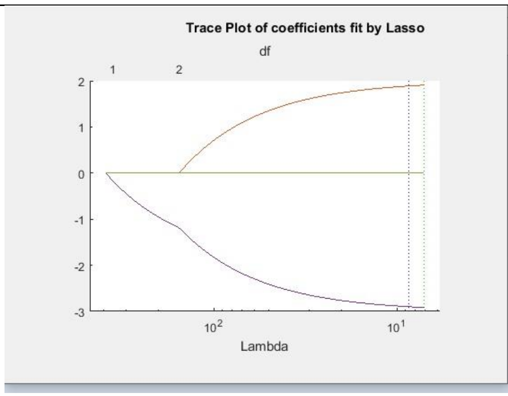
XN2 data



XN3 data



XN4 data



XN5 data

Fig(1) table showed that Lambda values in range $(10^0 - 10^1)$ for normal data, while they reached in range $(10^1 - 10^2)$ for all data with Outliers.

Table (1) Ans for Each Outlier Ratios

XN	XN1	XN2	XN3	XN4	XN5
0	-0.00131	0	0	0	0
1.885529	1.880074	1.909919	1.90644	1.899031	1.889054
0	0	0	0	0	0
-2.93665	-2.89604	-2.88625	-2.90192	-2.9095	-2.89923
0	0	0	0	0	0

Table (1) shows the ans values was similar for each normal and data with Outlier sets

Table (2) the P values (Min,Max and Average) Values for Each data sets

Data	Step	Min	Max	Average
XN	1	0	0	0
	2	0	1.885529466	1.022836529
	3	0	0	0
	4	-2.936653182	0	-2.230131267
	5	0	0	0
XN1	1	-0.020213675	0	-0.004349185
	2	0	1.890732496	0.862277094
	3	0	0	0
	4	-2.905287328	-1.38582E-15	-1.835865805
	5	0	0	0
XN2	1	0	0	0
	2	0	1.931842015	1.289433634
	3	0	0	0
	4	-2.91396312	0	-2.075751997
	5	0	0	0
XN3	1	0	0	0
	2	0	1.922314122	1.165941469
	3	0	0	0
	4	-2.918571598	0	-2.048316616
	5	0	0	0
XN4	1	0	0	0
	2	0	1.91616459	1.103649041

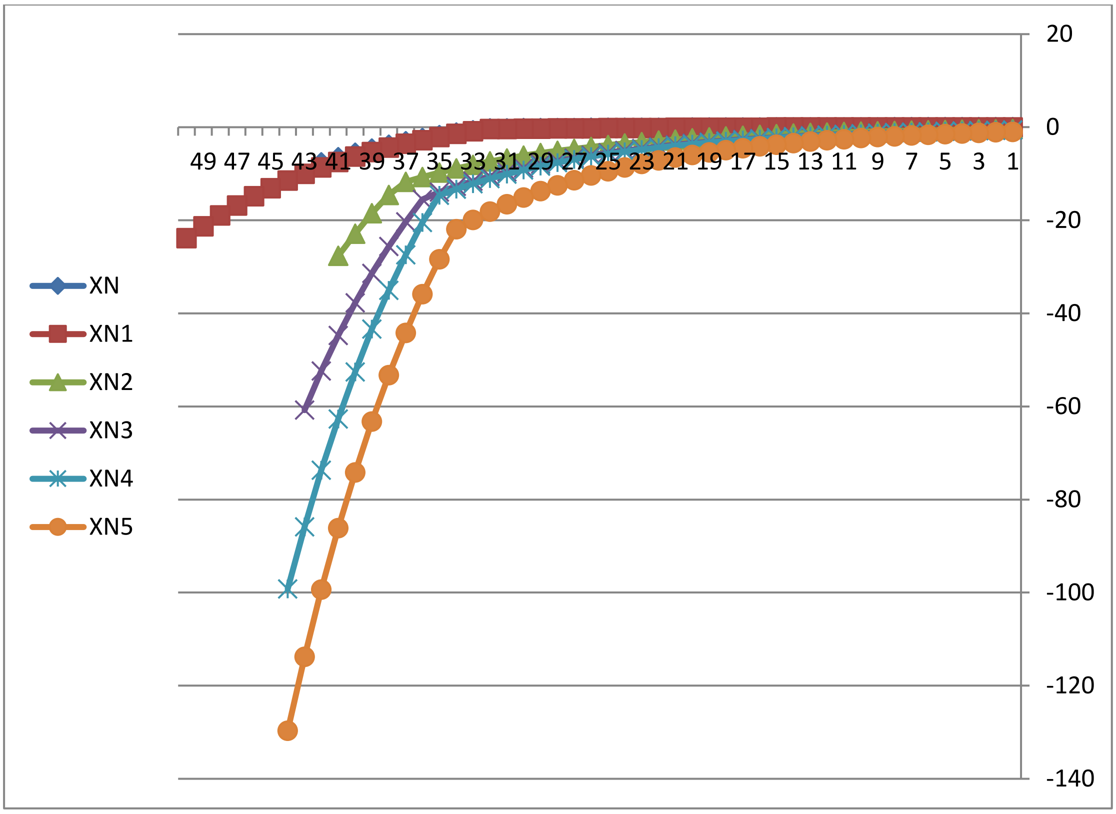
	3	0	0	0
	4	-2.924858312	0	-2.076941741
	5	0	0	0
XN5	1	0	0	0
	2	0	1.907887339	1.059404557
	3	0	0	0
	4	-2.916332166	0	-2.01891044
	5	0	0	0

Table (2) shows that P (Min Max and Average) values similar times and dissimilar other times for each normal and data with Outlier sets.

Table(3) The P values (Max,Min and Average) Values for (Fit intercept ,Fit Information of Lambda ,Mse and Se) for Each data sets .

		Min	Max	Average
Fit intercept	XN	-7.538622795	0.00411735	-0.928376614
	XN1	-23.8522127	-0.01428617	-3.548220074
	XN2	-27.621992	-0.400649023	-5.14744546
	XN3	-60.78486507	-0.579956104	-10.21472476
	XN4	-99.19880487	-0.608913872	-14.95873986
	XN5	-129.6729389	-1.009796502	-21.90383898
Fit information of lambda	XN	0.259579328	11.77160716	3.091551963
	XN1	1.350095845	128.8731905	28.73637854
	XN2	4.367466272	180.4642479	48.45385254
	XN3	5.324125388	264.9829752	68.09731129
	XN4	6.779522259	370.3162998	93.15786009
	XN5	7.134410742	389.7012926	98.03440603
Mse	XN	0.139788703	158.0361734	27.25233686
	XN1	34.68364894	23344.08728	6285.113166
	XN2	223.6426728	88418.83625	25177.57424
	XN3	268.2681393	132868.5317	35720.19782
	XN4	286.3038251	211600.7792	49690.49518
	XN5	335.5452633	240964.1312	57715.16864
Se	XN	0.020054	19.0571	3.570341
	XN1	15.5296	11352.37	3032.085
	XN2	178.7832	40576.47	15646.62
	XN3	189.5692	42613.55	17261.82
	XN4	160.165	51213.93	17334.63
	XN5	159.946	57628.53	18234.25

Table (3) Shows that (Mse) effected by Outlier ratios (Mse highly increase with increasing Outlier ratios)



Fig(2) fit-intercept values for Each data sets

Fig(2) Shows that fit-intercept values for Each data sets was decreasing in increasing Outlier rations.

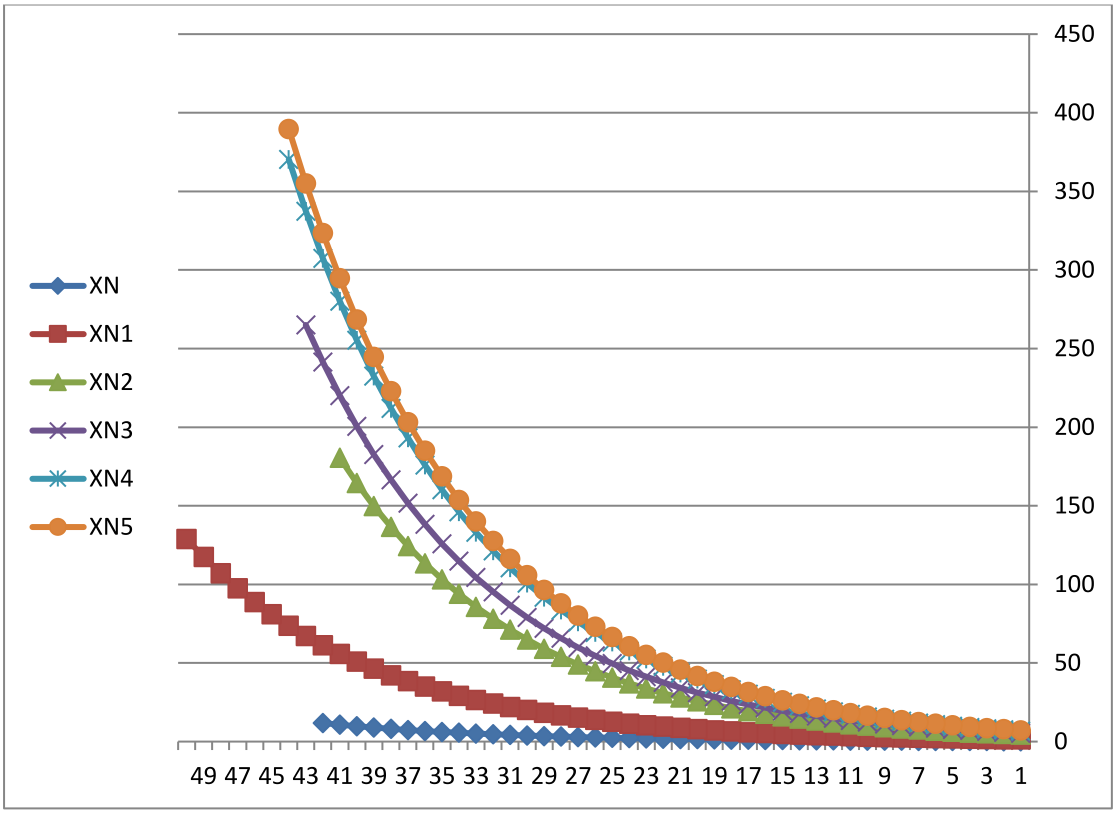
Table (4) the (Max,Min and Average) Values for (Lambda Min Mse and Lambda Max Mse) for Each data sets

	XN	XN1	XN2	XN3	XN4	XN5
Lambda Min Mse	0.259579	1.350096	4.367466	5.324125	6.779522	7.134411
Lambda Max Mse	0.259579	1.481729	5.773533	6.412928	8.165959	8.593423

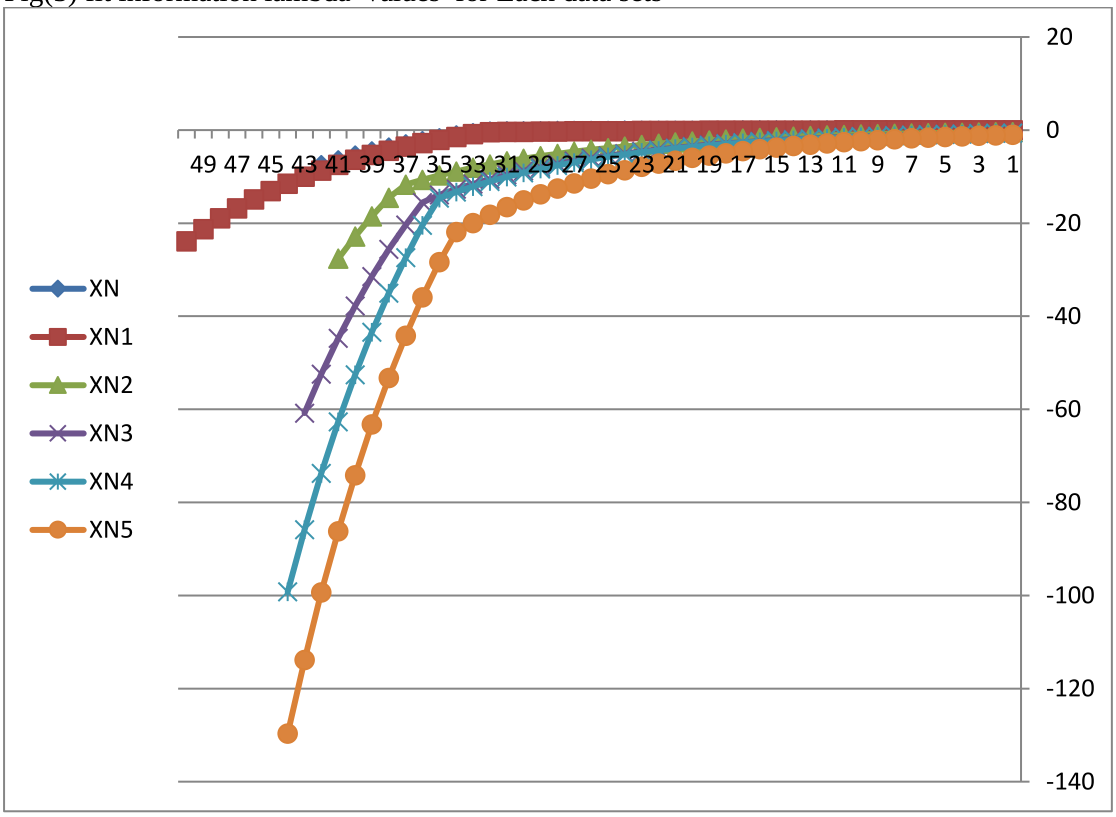
Table (4) shows that Lambda (Min and Max) values was effected and increasing by increasing Outlier ratios .

Table (5) Fit Information for Each data sets

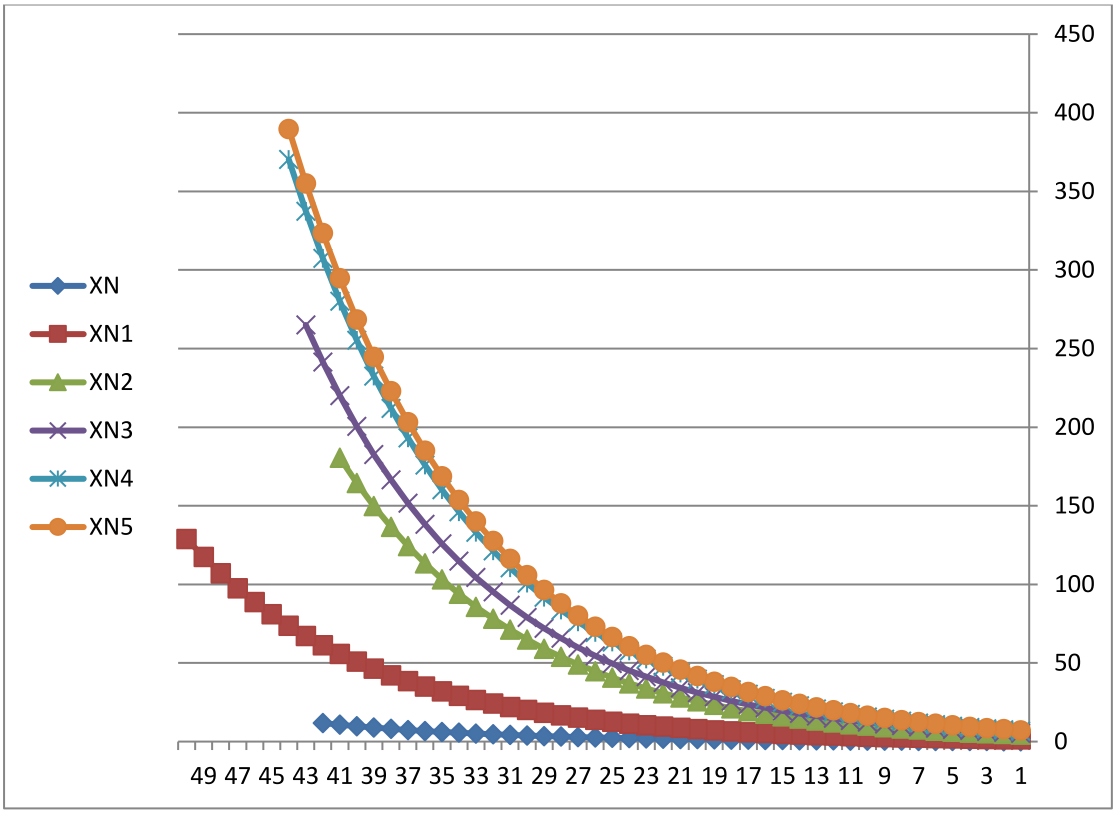
XN	XN1	XN2	XN3	XN4	XN5
-0.00384	-0.00047	-0.00039	-0.00021	-0.0001	-4.29E-05
1.995239	1.999485	1.999828	1.999939	1.999955	2.000016
0.001376	0.000694	0.000266	4.02E-05	1.24E-05	-8.87E-05
-2.99931	-3.00005	-3.00006	-2.99997	-2.99997	-2.99993
0.00315	0.000225	9.90E-05	2.84E-05	2.25E-05	-3.03E-06



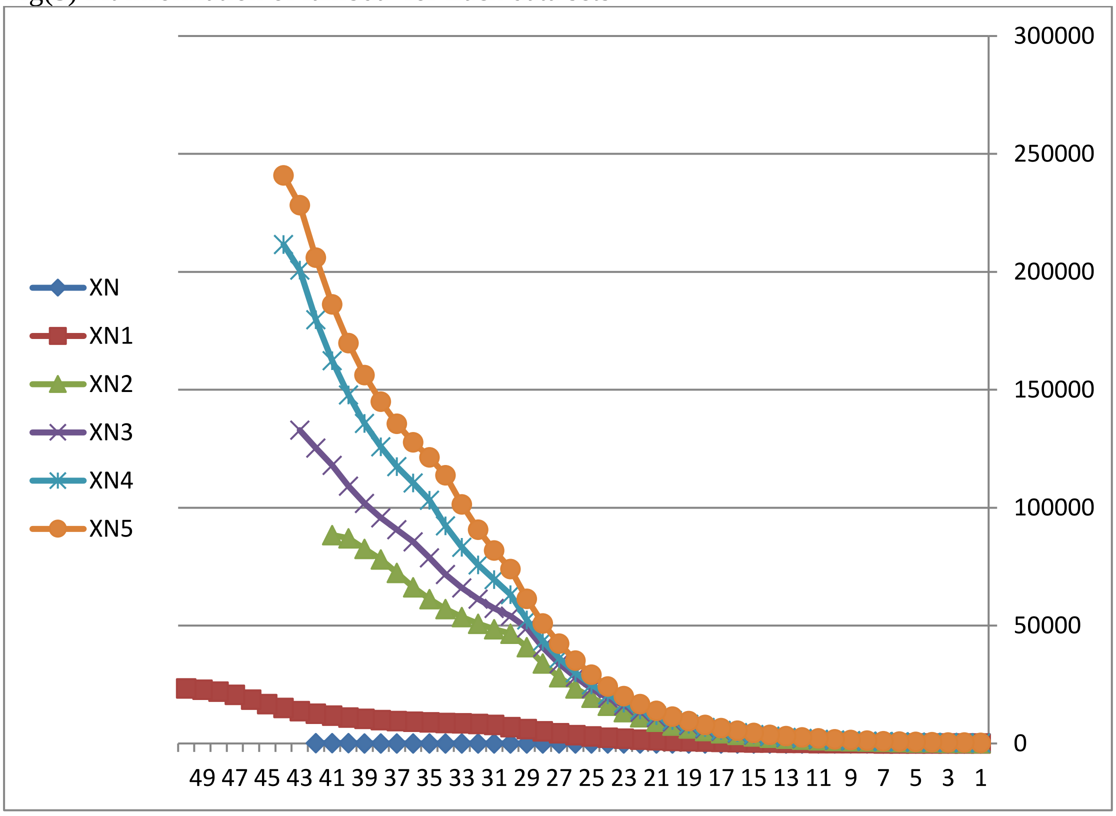
Fig(3) fit information lambda values for Each data sets



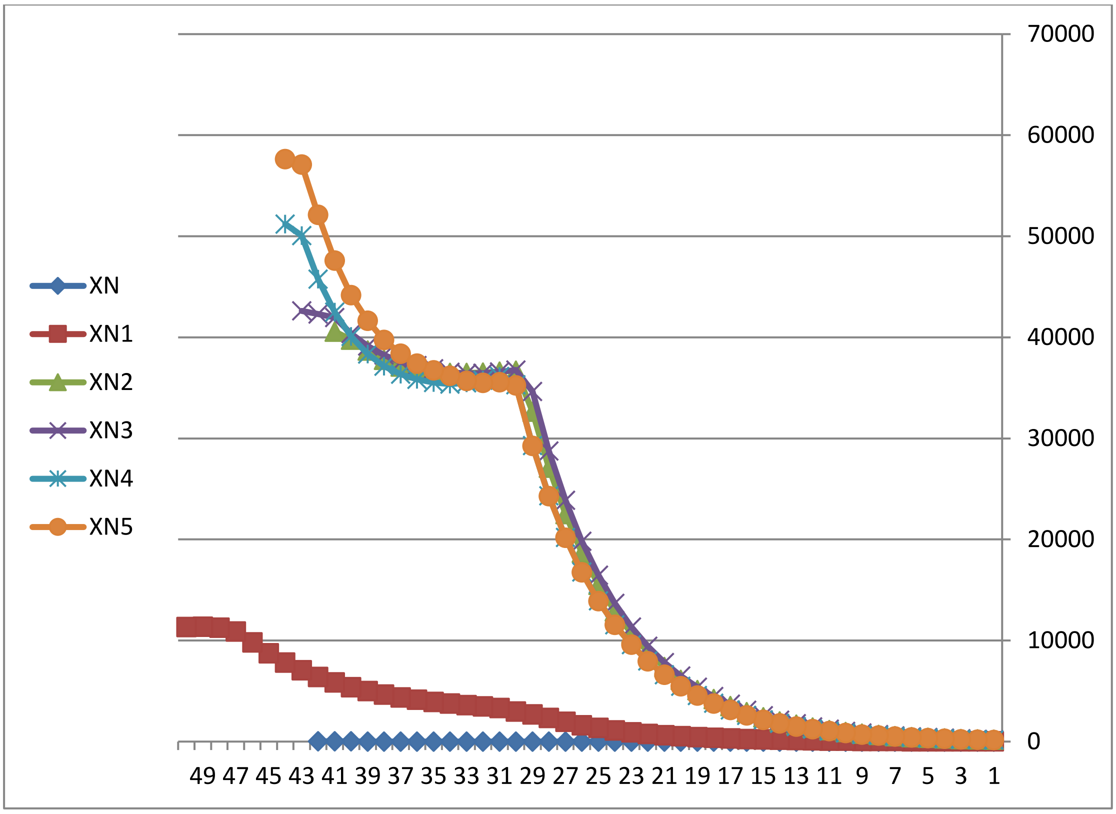
Fig(4) fit of intercept for Each data sets



Fig(5) Fit information of lambda for Each data sets



Fig(6) Mse values for Each data sets



Fig(7) Se values for Each data sets

Fig (3-7) shows that each values of them was effected increasing some times and decreasing other times with increasing Outlier ratios

Conclusions and Suggestions

1-Lasso regression estimators affected by Outlier ratios

2-Mean Square Errors for lasso regression effected by Outlier ratios

3-in increasing of Outlier ratios some of lasso parameters will increase and others will decrease

4-kernal regression can be compared with lasso regression in data include Outlier

5-other Outlier ratios can be included in the data sets

1. **Refrencess**

2. Ahrens A, Hansen CB, Schaffer ME. lassopack: Model selection and prediction with regularized regression in Stata. The Stata Journal. 2020;20(1):176-235.
3. Hans C. Bayesian lasso regression. Biometrika. 2009;96(4):835-45.
4. Hastie T, Taylor J, Tibshirani R, Walther G. Forward stagewise regression and the monotone lasso. Electronic Journal of Statistics. 2007;1:1-29.
5. Huang J, Ma S, Zhang C-H. Adaptive Lasso for sparse high-dimensional regression models. Statistica Sinica. 2008:1603-18.
6. Rahman SK, Sathik MM, Kannan KS. Multiple linear regression models in outlier detection. International Journal of Research in Computer Science. 2012;2(2):23.
7. Reid S, Tibshirani R, Friedman J. A study of error variance estimation in lasso regression. Statistica Sinica. 2016:35-67.
8. Roth V. The generalized LASSO. IEEE transactions on neural networks. 2004;15(1):16-28.
9. She Y, Owen AB. Outlier detection using nonconvex penalized regression. Journal of the American Statistical Association. 2011;106(494):626-39.
10. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267-88.
11. Alhayani, B. and Abdallah, A.A. (2020), "Manufacturing intelligent Corvus corone module for a secured two way image transmission under WSN", Engineering Computations, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/EC-02-2020-0107

12. H. S. Hasan, A. A. Abdallah, I. Khan, H. S. Alosman, A. Kolemen *et al.*, "Novel unilateral dental expander appliance (udex): a compound innovative materials," *Computers, Materials & Continua*, vol. 68, no.3, pp. 3499–3511, 2021. *doi:10.32604/cmc.2021.015968*
13. Alhayani, B., Abbas, S.T., Mohammed, H.J. et al. Intelligent Secured Two-Way Image Transmission Using Corvus Corone Module over WSN. Wireless Pers Commun (2021). https://doi.org/10.1007/s11277-021-08484-2