Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Estimating General Linear Regression model by Using Sure Independence Screening SIS Method under High Dimensional Data with Application

Mohammed Jassim Farhan, Ahmed Mahdi Salih

Statistics Department, College of Administration, and Economics, University of Wasit

Corresponding Author: Ahmed Mahdi Salih

Email: amahdi@uowasit.edu,iq , ahmadaljomaily@yahoo.com

Mobile: 07803658769

Article Info

Page Number: 4936 - 4943

Publication Issue:

Vol 71 No. 4 (2022)

Article History

Article Received*: 25 March 2022*

Revised*: 30 April 2022*

Accepted*: 15 June 2022*

Publication*: 19 August 2022*

Abstract

Estimating the general linear regression model when there is a high dimensions data understudy is very important topic to study and analyze, due the problems that appear when the dimension of data exceeds like multicollinearity, the study aims at using different adaptive penalized function to estimate the general linear regression model by using **Sure Independence Screening SIS**, in addition other method were chosen which is **Ridge Regression**. Data were collected that represent **Social Deprivation Index SDI** in Iraq from different governments of Iraq, moreover; simulation data were held and comparison were made to choose the better estimating method by using Mean Square Errors MSE.

Keywords: Sure Independence Screening, Panelized Functions, Ridge Regreesion

1- Introduction

Information about variables that we study is the key issue to determine a good methos to analyze. And when there is high dimension data under study the classical methods like Maximum Likelihood ML or the ordinary Least Squares OLS is not effective and weak.

In high dimension data Reducing dimensions of data under study has concerned by many scholars like ***Fan &Lv***[4] (2007) who introduced a new iterative selection variable procedure they called sure independence screening SIS. The researches depended on choosing initial estimation for the linear regression model than obtaining the loss function of the variables and selecting certain variables with the smallest values of the loss function. The next step is estimating the model that

contains the selected variables with one of the penalized regression methods then repeat that until we have the important variables. And ***Fan&Lv***[8] (2009) introduced a new penalized regression methods, they proposed a new rational ridge parameter depends on the ratio between L1-Norm and L2-Norm . Additionally, the researchers also used an initial estimation for linear high dimensional regression model parameters depends on the Pseudoinverse for the matrix $'X$. Simulation results showed that the new estimator is better than SCAD and Lasso, depending on the value of mean square errors.Lv& Fan called the new estimator smooth integration of counting and absolute deviation SICA.

In our study, we have a general linear regression model with several explanatory variables

$$\boldsymbol{Y = X\beta + \varepsilon} \qquad \textbf{... (1)}$$

Where $\boldsymbol{Y}$ represents $\boldsymbol{(n \times 1)}$dependent variable vector and $\boldsymbol{X}$ is $\boldsymbol{(n \times p)}$ explanatory where ($\boldsymbol{P > n)}$variables matrix containing a large number of variables, while $\boldsymbol{\varepsilon}$ is $\boldsymbol{(n \times 1)}$ random error vector and $\boldsymbol{\beta}$is (p × 1) parameters vector [10].

Regression models usually used in diverse statistical issues where there is a large number of variables and there is a small size sample under study many authors present various methods to estimate the parameter of the model in (1). High dimensions of the data may bring complex issues like noise accumulation and spurious correlation. Our study consists of 7 sections, section one is the introduction and section tow is Ridge Regression, while section three is Sure Independence Screening SIS Method, , section four is the concept of Social Deprivation Index SDI, section five is the criteria of comparison section six is the data and results and section seven is the conclusions and recommendations.

2- Ridge Regression

Estimating the general regression in (1) require an efficient estimation method due the condition of high dimensional data and Ridge regression is one the methods that suggested to analyze high dimensional data, it is a penalized regression which is simply a linear style to deal with high dimensional data [10].the Ridge estimation methods is simply can be written as a penalized regression minimizing errors subject to the additional condition that is called penalty function [9].

$$\widehat{\boldsymbol{\beta}}^{\boldsymbol{Ridge}} = \mathbf{arg\,min}_{\boldsymbol{\beta}} \frac{\mathbf{1}}{\boldsymbol{n}} (\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} + \boldsymbol{\lambda I}\|\boldsymbol{\beta}\|_{\mathbf{2}}) \qquad \textbf{.. (2)}$$

Where $\boldsymbol{I}$ is $(\boldsymbol{p} \times \boldsymbol{p})$identity matrix and $\|\boldsymbol{\beta}\|_{\boldsymbol{2}} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\beta}_{\boldsymbol{j}}^{\boldsymbol{2}}$ and by resolving the optimization in (2) and minimize the sum of square errors. Ridge regression has some very good qualities and it's good choice for high dimensions analysis. In terms of matrices, the optimization in (2) will be as follows [13].

$$\widehat{\boldsymbol{\beta}}^{\boldsymbol{Ridge}} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\lambda}\boldsymbol{I})^{-\boldsymbol{1}}\boldsymbol{X}'\boldsymbol{Y} \qquad \textbf{… (3)}$$

The estimation in (3) has been branded "Ridge Regression". Selecting the parameter λ which is called ridge parameterfascinates many scholars to present formulas depend on using L-norms or involve $\boldsymbol{\lambda}$ in quadratic forms the target is to minimize the errors In 2014 Dorugade [3] recommended a ridge parameter to use in case of high dimensional sets of data as follows.

$$\hat{\boldsymbol{\lambda}} = \frac{\boldsymbol{2\hat{\sigma}^2}}{\boldsymbol{K_{\max}}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \frac{\boldsymbol{1}}{\boldsymbol{\hat{\alpha}_j^2}} \qquad \textbf{… (4)}$$

Where $\boldsymbol{\hat{\sigma}^2} = \boldsymbol{Y}'[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}']\boldsymbol{Y}/(\boldsymbol{n} - \boldsymbol{p})$, and $\boldsymbol{K_{\max}}$the maximum value from Eigenvalues for the matrix $\boldsymbol{X}'\boldsymbol{X}$. And , $\boldsymbol{\alpha} = \boldsymbol{D}'\boldsymbol{\beta}$where $\boldsymbol{D}$is an orthogonal matrix, which implies $\boldsymbol{D}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{D} = \boldsymbol{\Lambda}$ where $\boldsymbol{\Lambda} = \boldsymbol{diag}(\boldsymbol{k_1}, \boldsymbol{k_2} \ldots \boldsymbol{k_p})$consist of the eigen values of $X'X$.

3- Sure Independence Screening SIS

Variable selection plays a vital part in high dimensional regression analysis, which means choosing the variables that have a excessive impact on the dependent variable. The penalized estimation methods are recognized on the choice of the penalty function, some of them choose easy and simple formula of the penalty function [11].

The method of Sure Independence Screening ***SIS*** is dissimilarto the penalized regression methods, since it uses the penalized methods to growa iterated method of selecting variables that depend on a pervious process to select the variables with marginal loss function, in general form the marginal loss function [5].

The first thing to do in SIS procedure is to have initial parameters estimation of the regression model in (1), some researchers advise using the following [14].

$$\boldsymbol{\theta} = (\boldsymbol{X}'\boldsymbol{X})^{+}\boldsymbol{X}'\boldsymbol{Y} \textbf{… (5)}$$

where + here signifies the Moore-Penrose inverse or Pseudoinverse [2], as an initial estimation. Now $\hat{\beta}_1$ represents the estimator in the first stage used to obtain the marginal utility vector, where $0 < \lambda < 1$as in (4), then we will use the parameters in (5) to obtain [6].

$$\boldsymbol{L_1 = \left(Y - \widehat{Y}\right)^2 / p} \quad \textbf{... (6)}$$

now we ranking L_1 from (6) from the small to large and we select the ***dj*** smallest values of it where d as follows and j represent the number of iteration[8].

$$\boldsymbol{d_1 = p/\log p} \quad \textbf{... (7)}$$

The d_1 variables selected from the marginal loss vector will be set to new model.

$$\boldsymbol{Y = X_{d1}\beta + \varepsilon} \quad \textbf{... (8)}$$

Where ***d1*** is the number of selected variables in the first step and we use anotherpenalized regression method to estimate the parameters of the model (8). In this point, we select the penalized least squares PLS to estimate the parameters of the new model by minimizing the following optimization [10].

$$\boldsymbol{\widehat{\beta}_1^P = \left(\left(X_{d_1}{}'X_{d_1}\right) + \eta I_d\right)^{-1} X_{d_1}{}'Y} \quad \textbf{... (9)}$$

Fan and Li [10] suggested the penalty function $\boldsymbol{\eta}$ as follows [7].

$$\boldsymbol{\eta = \frac{a\lambda - \|\theta\|}{(a-1)\lambda}} \quad \textbf{... (10)}$$

Where $\theta = (X'X)^{+}X'Y$, and $a = 3.7$was suggested by Fan and Runze, $\boldsymbol{\widehat{\beta}_1^P}$ is the penalized regression estimation so this parameters of (9) will be used to obtain the marginal loss as in (6).and so on we continuous until we have no variable to exclude. If we suppose thatt is the total number of variables in the last stage and the new regression model is $\boldsymbol{Y = X_t\beta + \varepsilon}$then the Sure Independent Screening estimator will be as follows [8].

$$\boldsymbol{\widehat{\beta}^{SIS} = (X_t{}'X_t)^{-1}X_t{}'Y} \quad \textbf{... (11)}$$

4- Social Deprivation Index SDI

Its very important to explain the concepts of the Social Deprivation Index SDI which consist of seven domains represent a ratios as in the following table

Table (1) Social Deprivation Index Domains

Domain	Variable
Income	Percent population having fair income (more than 200$ PM)
Education	Percent population 25 years or more with less than 12 years of education
Employment	Percent non-employed
Housing 1	Percent population living in renter occupied and crowded housing units
Housing 2	Percent population living in crowded housing units
Household Characteristics	Percent single-parent households with dependents < 18 years
Transportation	Percent population with no car

Here we have seven domains if we denote them by $\boldsymbol{x_1, x_2, \dots, x_7}$then the Social Deprivation Index SDI will be as follows [1].

$$\boldsymbol{SDI = \sum_{i=1}^{7} x_i \dots} \textbf{ (12)}$$

If SDI is more than 2 then the household suffers from depravation.

5- Criteria of comparison

Comparison among estimators is a key process in any statistical or scientific research that could help the researchers to determine the best statistical method of analyzing or model selection [4].

There are a variety of statistical comparison measures that are founded on a certain assumption or theoretical foundation. We have chosen Mean Square Errors MSE as follows.

$$\boldsymbol{MSE = \sum \left(Y_i - \hat{Y}_i\right)^2 / n} \qquad \textbf{... (13)}$$

6- Data and Results

We have collected huge number of sets of surveys data from the Central Statistical Organization IRAQ to represent 300 group of families from variety parts of the country, and

we calculate the SDI vector (300×1), we have 300 variables from many topics like health, education and living standards etc.

Here we subdivide our 300 group of families (***n=300***) to start with 20 until the sample size will be equal to the number of variables, we obtain for Ridge regression RR estimators as in (3) and Sure Independence Screening SIS as in (11) and finally we calculate **MSE** for both methods as in (13) in the following table (2) By using MATLAB as follows.

Table (2) MSE for the Estimation Methods (RR, SIS)

n	RR	SIS	n	RR	SIS	n	RR	SIS
20	**99.2923**	**78.3624**	**140**	**70.7582**	**52.95065**	**260**	**41.9815**	**37.09717**
40	**96.7018**	**75.06633**	**160**	**66.0844**	**50.14106**	**280**	**36.3915**	**36.91072**
60	**88.1691**	**69.65052**	**180**	**62.8386**	**47.72262**	**300**	**30.3279**	**35.95859**
80	**83.9365**	**64.71824**	**200**	**52.8296**	**41.67422**			
100	**80.6698**	**60.49849**	**220**	**50.7313**	**39.82404**			
120	**72.4651**	**54.01582**	**240**	**47.4806**	**38.44945**			

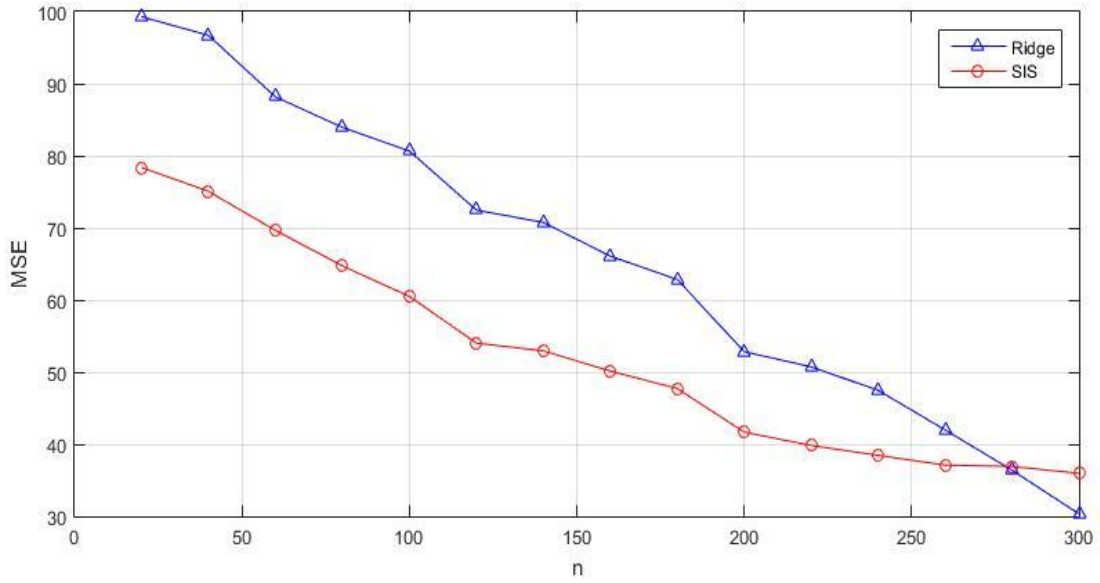

Figure (1) MSE (RR, SIS)

7- Discussion

From Table (1) and Figure (1) we can see that when the sample size is very small compared to the number of variables the SIS estimator perform significantly better than RR estimator and the situation still the same until the sample size go very close to the number of variables then the RR estimator becomes better than SIS estimator. Therefore, we recommend to use SIS estimator when the sample size is very small compared to the number of variables.

References

[1]. Acharjya. D, Kauser. A, "A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools" *International Journal of Advanced Computer Science and Applications.* Vol. 7, No 2, pp. 511-518, 2016.

[2]. Courrieu. P, "Fast Computation for Moore-Penrose Inverse Matrices" Neural Information Processes – Letters Reviews, Vol. 8, No. 2, pp. 25-29, 2005.

[3].Dorugade. A, "New Ridge Parameters for Ridge Regression" Journal of the Arab Universities for Basic and Applied Statistics, Vol. 15, No. 3, pp. 94-99, 2014.

[4].Fan. Y, Lv. J, "Sure Independence Screening for Ultra-High Dimensional Feature Space" Royal Statistical Society, Vol. 70, Issue. 5, pp. 849-911, 2008.

[5]. Fan. J, Samworth. S, Wu. Y, "Ultrahigh Dimensional Feature Selection: Beyond the Linear Model" Journal of Machine Learning Research, Vol. 10, pp. 2013-2038, 2009.

[6]. Fan. J, Runze. L, "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties" Journal of American Statistical Association, Vol.96, pp. 1348-1360, 2001.

[7]. Fan. J. Li. R, "Statistical Challenges with High Dimensionality: Features Selection in Knowledge Discovery" proceedings of International Congress of Mathematics, Vol. 3, pp. 595-622, 2006.

[8]. Fan. Y, Lv. J, "A Unified Approach to Model Selection and Sparse Recovery Using Regularized Least Squares" *The Annals of Statistics*, Vol. 37, No. 6A, pp. 3498-3528, 2009.

[9]. Fan. Y, Lv. J, "Sure Independence Screening for Ultra-High Dimensional Feature Space" Royal Statistical Society, Vol. 70, Issue. 5, pp. 849-911, 2008.

[10].Hoerl. A, Kennard. R, "Ridge Regression: Biased Estimation for Nonorthogonal Problems" *Technometrics*, Vol. 12, No. 1, pp. 55-67, 1970.

[11]. Hastie. T, Tibshirani. R, Friedman. J, "*The Elements of Statistical Learning: Data Mining, Inference, and Prediction*" Springer Series in Statistics, USA, 2010.

[12].Kapetanios. G, Marcellino. M, Petrova. K, "Analysis of the Most Recent Modeling Techniques for Big Data with Particular Attention to Bayesian

[13].Mohammed. L, Khadhm. S, "Estimate Kernel Ridge Regression Function in Multiple Regression" *Journal of Economics and administrative Science*, Vol. 24, No. 103, pp. 411-419, 2018.

[14]. Salih. A, Hmood. M, "Analyzing big data sets by using different panelized regression methods with application: surveys of multidimensional poverty in Iraq" Periodicals of Engineering and Natural Sciences. Vol. 8, No. 2, pp. 991-999, 2020.