IRAQI STATISTICIANS JOURNAL

https://isj.edu.iq/index.php/isj

ISSN: 3007-1658 (Online)

Estimating General Linear Regression Model of Big Data by Using Multiple Test Technique

Ahmed Mahdi Salih[1], Munaf Yousif Hmood[2]

[1]Department of Statistics / College of Administration and Economic, University of Wasit / Iraq
[2]Department of Statistics / College of Administration and Economics, University of Baghdad / Iraq

ARTICLE INFO

Article history:

Received	13/11/2024
Revised	16/11/2024
Accepted	14/1/2025
Available online	15/5/2025

Keywords:
Big Data
Ridge regression
Social Deprivation
Multiple tests

ABSTRACT

Big Data analyses attract many researchers to create or develop new efficient statistical techniques to analyse big sets of data and deal with the problems that Big Data bring like noise accumulation and multicollinearity. This work presents an innovative approach to estimate the generic linear regression model of Big Data using several test processes. Researchers are faced with a great problem when it comes to big data analysis, which is why they should be developing new techniques for estimating the general linear regression model. Information has been collected from the Central Statistics Organization IRAQ which is represented by the Social Deprivation Index SDI. Where the concept of the SDI indicator was cleared, and all its contents were, and we showed how the SDI indicator was calculated. Two methods have been chosen to estimate the general linear regression model: our proposed method, which represents an adapted estimation method of the OCMT estimation method by using a ratio of quadratic forms as a multiple test procedure to select the variables in the general linear regression model, and the traditional method, Ridge regression RR, which is present to deal with big sets of data. One measure that has been used to compare the approaches is the mean square error, or MSE. Here we compare one classical method RR which depends on adding some positive quantities to avoid singularity of X'X matrix and a proposed method that depends on selecting variables. Last, we conclude that our proposed estimator, which depends on the multiple test procedure, is the best and has the best performance.

1. Introduction

Analyzing Big Data sets becomes a very substantial issue and a leading topic for many studies and searches, it gives a great challenge for researchers and data analyzers to find or develop new and efficient analyzing methods to solve the problems that appear in data when data dimensions get larger, researchers offered definitions for Big Data such like Bühlmann [1]

Boyd and Crawford [2] elucidated it as “A cultural, technological, and scholarly phenomenon that rests on the interplay of Technology, Analysis and Mythology.”

Chang [3] and Chudik [4] define it as “The data sets and analytical techniques in applications that are so large and complex that they require advanced and unique data storage, management, analysis, and visualization technologies”

Researchers used algorithm schemes to define their methods and approaches for analyzing data that should be understandable. Big data is therefore commonly used in scientific domains

[*] Corresponding author. E-**mail** address: munaf.yousif@coadec.uobghdad.edu.iq

https://doi.org/10.62933/e6740m35

including marketing, healthcare, demography, and several other areas.

Numerous scholars have focused their attention and primary concern on big data concerns. It pushed them to introduce fresh statistical approaches and methods because of how quickly life and technology are developing across all domains. [5]. As a result, many researchers have investigated Big Data, and the selected authors are listed below.

Hoerl and Kennard [6] investigated the performance of Ridge regression by using different Ridge parameters. By choosing the biasing parameter in Ridge regression, they first thoroughly compared the maximum likelihood estimator and the estimation for Ridge regression. They also concluded that the chance that Ridge regression produces fewer square errors than MLE is more than 0.5 and that the mean square error of the regression coefficients in Ridge is lower than in MLE.

To choose effective variables with high dimension criteria for multiple testing of the explanatory variables, Benjamini and Hochberg [7] proposed a novel method for regulating the false discovery rate (FDR) They offered a different method to deal with multiple testing problems, a simulation was made to compare the new method with the Bonferroni-Type procedure for controlling false discovery. They proved their preference for the new method, as they supported their study with many numerical examples.

Pesaran and Smith [8] presented the least angle regression LARS as a selection procedure when there are many variables in the linear regression model. They derive the theoretical properties of their new procedure, LARS procedure uses the LASSO estimator with some improvement to select the effective covariates from a large number of covariates, simulation results showed the efficiency of the new LARS selection procedure.

Barrientos and Peña [9] introduce new data-subletting algorithms to approximate and scale the implementation of the Bayesian bootstrap in massive Big Data Sets of data and compare the new algorithm with two existing algorithms. Moreover, they derive the new algorithm's theoretical and computational properties and find out that the new algorithm provides a good strategy for the loss function of some class Bayesian bootstraps.

Velten and Huber [10] suggested a Bayesian tool procedure to provide a variable selection procedure for a linear regression model under high dimensions conditions. This procedure adapted to the regular penalized regression method by using variational Bayes. Simulation results with an extreme number of variables supported the recommendation of their new variable selection method.

2. Ridge Regression

To estimate a general regression model, one needs a robust estimation technique that can handle any kind of data issue. Traditional techniques like maximum likelihood and least squares require certain assumptions such as $\boldsymbol{E(\varepsilon_i) = 0}$ and $\boldsymbol{E\left(\varepsilon_i^2\right) = \sigma^2}$ that are hard to attend to in actual data.

Ridge regression is one of the early methods recommended to analyse large sets of data, it is a kind of penalized regression which is simply a linear approach to deal with large sets of data [11].

$$\hat{\beta}^{OLS} = (X'X)^{-1}X'Y \qquad (1)$$

The OLS estimators are efficient if the correlation formula of $X'X$ matrix is nearly a unit matrix, but if it is not nearly a unit matrix, then the estimators of OLS are not efficient because they are sensitive to the errors [12]. Furthermore, there is an extra requirement known as the penalty function that applies to the penalized regression minimizing errors.

$$\hat{\beta}^{Ridge} = arg\,min_{\beta}\frac{1}{n}(\varepsilon'\varepsilon + \lambda I\|\beta\|_2) \qquad (2)$$

Where I am $(p \times p)$ identity matrix and $\|\beta\|_2 = \sum_{j=1}^{p}\beta_j^2$. We will apply shrinkage over β, which minimizes the sum of square errors, by solving the optimization in (2). For high dimensions and Big Data analysis, ridge regression is a good option because it possesses a number of excellent features. In terms of matrices, the optimization in (2) will look like this.

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

The estimate in (3) has been called as "Ridge Regression." Selecting the complexity parameter λ, is known as the ridge parameter, it has been developed in many methods. Some of them recommend choosing value $0 \leq \lambda \leq 1$ On the other hand, the others suggest using λ which lowers the model's parameter mean square error.

Thus, if we use the conventional form of expression and assume that D is an orthogonal matrix, , which implies $D'X'XD = \Lambda$where $\Lambda = diag(\, k_1, k_2 \ldots k_p)$ involve of the eigen values of $X'X$; so the equivalent regression model will be

$$Y = X^*\alpha + \varepsilon \qquad (4)$$

where $X^* = XD$, $\alpha = D'\beta$, here $\hat{\alpha} = \Lambda^{-1}X^*Y$ symbolize the OLS estimators for the equivalent model [13].

Choosing the ridge parameter $\boldsymbol{\lambda}$ attracts many researchers to propose formulas that depend on using L-norms or involve $\boldsymbol{\lambda}$ in quadratic forms; the target is to minimize the errors.

In 2014, Dorugade proposed the following new ridge parameter to be used when dealing with huge data sets:

$$\hat{\lambda} = \frac{2\hat{\sigma}^2}{K_{\max}}\sum_{j=1}^{p}\frac{1}{\hat{\alpha}_j^2} \qquad (5)$$

Where $\hat{\sigma}^2 = Y'[I - X(X'X)^{-1}X']Y/(n-p)$, and $K_{\max}$ the maximum value from Eigenvalues for the matrix $X'X$.

3. Proposed Estimator

An important method for handling high dimensional data and Big Data sets is the multiple test process, which uses a test of the ratio of quadratic forms in normal variables as a selection test for the statistically relevant covariate. When [14] originally submitted the test, they added a score test based on the ratio of quadratic forms, and this test is not degenerate in large dimension situations. Therefore, using it with Big Data sets is appropriate [15]). To choose the covariates that have an impact on the dependent variable throughout the multiple testing process, we first define the ratio of the quadratic form test. Supposing there is a linear regression model , and we wish to test the hypothesis , Geoman begins with the following quadratic form test in order to obtain a test statistic under high-dimension conditions [16].

$$Q = n^{-1}(Y - X\beta)'ZZ'(Y - X\beta) \qquad (6)$$

Z is the standardized matrix of X, and because of the huge sample size, the division by n helps to prevent degeneracy (James & Stein, 1992) Beyond the power of testing for the test statistic in (6) is the quadratic form that follows the Chi-square distribution with r degree of freedom. The following divisions apply to the nuisance parameter test.

$$Q = \frac{(Y-X\beta)'ZZ'(Y-X\beta)}{n\sigma^2} \qquad (7)$$

Geoman recommended adopting a pivot approximation for the test statistic in (7) to be appropriate in the linear model scenario, given the null hypothesis and the high sample size as $\boldsymbol{n}\to\infty$

$$S = \frac{Q}{E(Q)} = \frac{(Y-X\beta)'ZZ'(Y-X\beta)}{trace\,(Z'WZ)} \qquad (8)$$

Since and the test statistic in (8) depend on attaining here may be issues with singularity for the matrix since big data sets are being studied. As a result, the test statistic's denominator in (8) can be stated as follows.

$$trace\,(Z'WZ) = \,(Y - X\beta)'D(Y - X\beta) \qquad (9)$$

By substituting (8) in (9), we can get [17].

$$S = \frac{(Y-X\beta)'ZZ'(Y-X\beta)}{(Y-X\beta)'D(Y-X\beta)} \qquad (10)$$

The test statistic in (10) is appropriate for the Big Data condition since we avoid obtaining Σ where D is the diagonal matrix of Σ . Lastly, the test statistic will be as follows after replacing the model parameters with the estimated values [18].

$$S = \frac{(Y - X\hat{\beta})' ZZ' (Y - X\hat{\beta})}{(Y - X\hat{\beta})' D (Y - X\hat{\beta})} \qquad (11)$$

As a ratio between two quadratic forms with a specified degree of freedom and a significant level, the test statistic in (11) has a F distribution, and we reject. if the equation on (11) will be employed in the multiple test procedure for choosing covariates that can affect over Y.

If we get a general linear model of the following form [19].

$$Y = X\psi + \varepsilon \qquad (12)$$

[8] originated an estimate for the parameter by means of asymptotic properties like the following:

$$\hat{\psi}_j = \gamma_j \left(X_j' M X_j\right)^{-1} X_j' M Y \qquad (13)$$

Where $\psi = \theta_j / \sigma_{jj}$ and $\theta_j = \sum_{h=k1+1}^{p} \beta_h$ and we can test the marginal and net impact of each covariate as the new parameter consider them on the model (12) as the first stage in our multiple testing procedure for testing the hypothesis $H_0: \psi_j = 0 \ \ Vs \ \ H_1: \psi_j \neq 0$ as follows [20].

$$S_j = \frac{(Y - X_j\hat{\psi}_j)' Z_j Z_j' (Y - X_j\hat{\psi}_j)}{(Y - X_j\hat{\psi}_j)' D (Y - X_j\hat{\psi}_j)} \qquad (14)$$

Assuming that at the conclusion of the first step is the matrix containing all of the statistically significant covariates, and is the matrix containing the remaining that had no significance and had not been selected in the first stage [21]

As we have no covariates that are statistically significant to be added. Our suggested estimate, the ordinary least square estimator, will be applied to the matrix , which at the end of the last stage contains all the important covariates [22].

$$\hat{\beta}^{PE} = \begin{cases} \hat{\beta}^{OLS} & (\beta \neq 0) \\ 0 & Otherwise \end{cases} \qquad (15)$$

Many statistical comparison techniques exist, each of which is predicated on a different premise or theoretical framework [23] . The mean square errors, or MSE, are the choices we have made.

$$MSE = \sum_{i=0}^{n} \left(Y_i - \hat{Y}_i\right)^2 / n - p \qquad (16)$$

4. Results and Discussion

Prior to the discussion of data details, a brief explanation of the Social Deprivation Index SDI was introduced and calculated for a group of families in the following table.

Table 1 Social Deprivation Index Domains

Domain	Variable
Income	Percent population having the 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 [19].

$$SDI = \sum_{i=1}^{7} x_i \qquad (17)$$

If SDI is more than 2 then the household suffers from deprivation. As a result, we have obtained numerous survey data sets from the Central Statistical Organization (IRAQ) to represent 10,000 family groups from different regions of the nation. We then compute the SDI vector (10000×1), which consists of 300 variables of different kinds, including quantitative, ordinal, nominal, and so on.

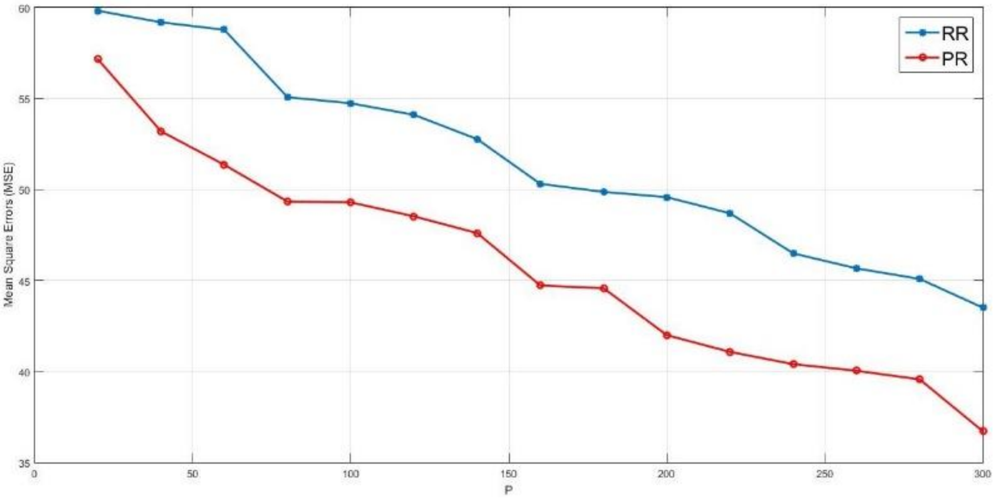

Figure (1) MSE (RR, PR)

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

P	*RR*	*PR*
20	59.82293	57.15657
40	59.18722	53.19552
60	58.77722	51.35891
80	55.07427	49.33837
100	54.73758	49.29392
120	54.10372	48.52413
140	52.76929	47.60551
160	50.30942	44.72516
180	49.86881	44.57246
200	49.57797	41.99591
220	48.69005	41.07331
240	46.48855	40.40143
260	45.66452	40.03867
280	45.08354	39.56356
300	43.50412	36.71205

We use real data to clarify all tables based on the number of variables p, adding 20 actual variables at a time to calculate the MSE for all estimators. The PR estimator performs well according to Table (1), but the **RR** estimator performs poorly. **PR** is the most accurate estimator for estimating the coefficients of the linear regression model, according to the findings of the real data analysis.

5. Conclusions

The suggested estimator, PR estimators perform best when real data is used and there are comparatively few variables. The suggested estimator PR approaches that rely on the multiple testing process to determine which variable is statistically significant are the best when considering Big Data features.

We suggest using our suggested estimator for estimating a linear regression model's coefficient in big data scenarios. The suggested estimator PR performs significantly well when real data is used, and it obtains the lowest Mean Square Errors (MSE) values for both small and large p values.

References

[1] P. Bühlmann and S. Van De Geer, Statistics for high-dimensional data: methods, theory and applications, Springer Science & Business Media, 2011.

[2] D. Boyd and K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," *Information, communication & society,* vol. 15, p. 662–679, 2012.

[3] W. L. Chang, N. Grady and others, "NIST big data interoperability framework: volume 1, big data definitions," 2015.

[4] A. Chudik, G. Kapetanios and M. H. Pesaran, "A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models," *Econometrica,* vol. 86, p. 1479–1512, 2018.

[5] A. De Mauro, M. Greco and M. Grimaldi, "What is big data? A consensual definition and a review of key research topics," in *AIP conference proceedings,* 2015.

[6] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics,* vol. 12, p. 55–67, February 1970.

[7] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological),* vol. 57, p. 289–300, 1995.

[8] M. H. Pesaran and R. P. Smith, "Signs of impact effects in time series regression models," *Economics Letters,* vol. 122, p. 150–153, 2014.

[9] A. F. Barrientos and V. Peña, "Bayesian bootstraps for massive data," 2020.

[10] B. Velten and W. Huber, "Adaptive penalization in high-dimensional regression and classification with external covariates using variational Bayes," *Biostatistics,* vol. 22, p. 348–364, 2021.

[11] J. Fan and J. Lv, "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society Series B: Statistical Methodology,* vol. 70, p. 849–911, October 2008.

[12] M. Nikolova, "Local Strong Homogeneity of a Regularized Estimator," *SIAM Journal on Applied Mathematics,* vol. 61, p. 633–658, January 2000.

[13] سهيل نجم عبود and ايناس صلاح خورشيد, "Comparison between the Methods of Ridge Regression and Liu Type to Estimate the Parameters of the Negative Binomial Regression Model Under Multicollinearity Problem by Using Simulation," *journal of Economics And Administrative Sciences,* vol. 24, 2018.

[14] J. J. Goeman, H. C. van Houwelingen and L. Finos, "Testing against a high-dimensional alternative in the generalized linear model: asymptotic type I error control," *Biometrika,* vol. 98, p. 381–390, May 2011.

[15] A. W. v. d. Vaart, Asymptotic Statistics, Cambridge University Press, 1998.

[16] J. J. Goeman, S. A. Van De Geer and H. C. Van Houwelingen, "Testing Against a High Dimensional Alternative," *Journal of the Royal Statistical Society Series B: Statistical Methodology,* vol. 68, p. 477–493, April 2006.

[17] A. M. Salih and M. Y. Hmood, "Analyzing big data sets by using different panelized regression methods with application: surveys of multidimensional poverty in Iraq," *Periodicals of Engineering and Natural Sciences (PEN),* vol. 8, p. 991–999, 2020.

[18] A. M. Salih and M. Y. Hmood, "Big data analysis by using one covariate at a time multiple testing (OCMT) method: Early school dropout in Iraq," *International Journal of Nonlinear Analysis and Applications,* vol. 12, p. 931–938, 2021.

[19] P. D. and K. Ahmed, "A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools," *International Journal of Advanced Computer Science and Applications,* vol. 7, 2016.

[20] K. Xu, "A new nonparametric test for high-dimensional regression coefficients," *Journal of Statistical Computation and Simulation,* vol. 87, p. 855–869, September 2016.

[21] R. Kumar, B. Moseley, S. Vassilvitskii and A. Vattani, "Fast Greedy Algorithms in MapReduce

and Streaming," *ACM Transactions on Parallel Computing,* vol. 2, p. 1–22, September 2015.

[22] W. James and C. Stein, "Estimation with Quadratic Loss," in *Breakthroughs in Statistics,* Springer New York, 1992, p. 443–460.

[23] لقاء علي محمد and صابرين حسين كاظم, "Estimate Kernel Ridge Regression Function in Multiple Regression," *journal of Economics And Administrative Sciences,* vol. 24, 2018.