Journal of Economics and Administrative Sciences (JEAS)

Available online at http://jeasiq.uobaghdad.edu.iq

Using Quadratic Form Ratio Multiple Test to Estimate Linear Regression Model Parameters in Big Data with Application: Child Labor in Iraq

Assistant Prof. Ahmed Mahdi Salih
Dept. of Statistics. College of Administration and Economics
University of Wasit

Prof. Dr. Munaf Yousif Hmood
Dept. of Statistics. College of Administration and Economics University of Baghdad

Received:18/8/2021 **Accepted: 22/8/2021** **Published: March / 2022**

Abstract:

The current paper proposes a new estimator for the linear regression model parameters under Big Data circumstances. From the diversity of Big Data variables comes many challenges that can be interesting to the researchers who try their best to find new and novel methods to estimate the parameters of linear regression model. Data has been collected by Central Statistical Organization IRAQ, and the child labor in Iraq has been chosen as data. Child labor is the most vital phenomena that both the society and education are suffering from and it affects the future of our next generation. Two methods have been selected to estimate the parameter of linear regression model, one Covariate at a Time Multiple Testing OCMT. Moreover, the Euclidian Distance has been used as a comparison criterion among the three methods.
Paper type Research paper.

Keywords: **Big Data, OCMT, Multidimensional Poverty, Child Labor . OCMT**

1- Introduction:

Big Data, as a term, has been developed lately in numerous topics that have a relation to academic and industrial issues [14]. Big Data Analysis becomes a main focus due to the fast growing of Big Data and the huge size of data that needs analysis such as internet profiles, satellite images, personal information update, etc…; the key idea of Big Data is gigantic data taken from diverse sources in variety of types. Many researchers presented definitions for Big Data terms according to the dimensions terms of it such as:

Schroeck et al. [20] defined it to be "Big Data is a combination of Volume, Variety, Velocity and Veracity that creates an opportunity for organizations to gain competitive advantage in today's digitized marketplace."

Chang and Grady [4] described Big Data as "Extensive datasets, primarily in the characteristics of volume, velocity and/or variety that require a scalable architecture for efficient storage, manipulation, and analysis."

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

In order to express their methods and techniques in analyzing data that should be easy to understand, researchers adopted algorithms schemes. Big Data appears in frequent fields of science, like marketing, healthcare, demography, and several other fields.

Big Data's concerns and challenges have been the main concern and the attention of many researchers and it urged them to present new statistical methods and techniques due to the fast development of technology and life at all fields. Big Data has been investigated by many researchers, selected authors are listed below.

Hoerl & Kennard [10] (1970) projected an innovative estimator for linear regression model in case of a large set of data under study. They studied correlation matrix of explanatory variables under high dimensions. The researchers observed that this correlation matrix will not be close to unit matrix as an assumption of linear regression, so they recommended adding positive amounts to be added to the diagonal of $X'X$ in order to avoid singularity and unsatisfied sum of the minimum square of residuals. Furthermore, they succeeded in labeling the new estimator with Ridge Regression.

Nikolova [16] (2000) examined a variety of penalized regression methods, using diverse penalty functions. His study considers L1-Norm and L2-Norm. She suggested the usage of a ratio of two norms so as to develop the penalty function under extreme high dimensional conditions. The simulation's results reinforced the new mixture penalty function over the usual penalty function that uses one L-Norm. Nikolova studied the Bayesian estimators via using a mixture of penalty function . He, also, used the linear regression model estimators.

Fan &Lv [6] (2007) introduced a new iterative collection variable procedure, they are called sure independence screening SIS. The researches relied on the choice of initial estimation for the linear regression model rather than attaining the loss function of the variables and selecting certain variables with the smallest values of the loss function.

The subsequent step is estimating the model that comprises the selected variables with one of the penalized regression methods, then this will be repeated until we have the significant variables. The researchers compared the new estimators with Dantzig selector and the adoptive lasso and the result revealed that SIS has a better performance under high dimensions conditions.

Pesaran & Smith [17] (2014) introduced a novel class of estimation for the regression coefficients, depending on the overall and the marginal effect of the coefficients themselves. They investigated correlation among regressors in time series models and regression models under high dimensions data. The new estimators are utilized in variable selection methods to determine which covariates have a marginal and overall effect on the dependent variable. The new estimators' utility in Big Data analysis was confirmed by simulation results.

Chudik et al. [5] (2018) offered a variety of approaches for estimating the linear regression parameter, with the new method being iterative and based on evaluating each covariate one at a time. They used the overall and marginal impact estimations were submitted by Pesaran and Smith [17], and they suggested an adaptive test for each covariate under Big Data conditions. They created a simulation study for the comparison of their estimator with other estimators such as Lasso and SCAD. Results exhibited that the new estimator is better under the quadratic loss function. The researchers called their method with a one-covariate at a time multiple testing OCMT.

2- Big Data Analysis Based on Greedy Algorithms

Big Data analyzers develop new statistical approaches to analyze the massive amount of data under study, many researchers employed new iterative methods to analyze data they named it Greedy Algorithms and; they defined it as [13].

"One of the simplest algorithms to implement: take the closest/nearest/most optimal option, and repeat. It always chooses which element of a set seems to be the best at the moment. It never changes its mind at a later point."

Greedy algorithm work from top to down, that is to say: the algorithm makes one selection and then another one, subsiding problems to minor ones.

3- One Covariate at a Time Multiple Testing OCMT

OCMT method was presented in 2016 by Chudik, Kapetanios and Pesaran [13] they suggested a new model section procedure for Big Data sets. The chief idea is to test every regression coefficient β_j individually one by one with a focus on the marginal and net impact of regressors in the first step, then this procedure is iterated until all statistically important covariates are involved in multiple regression. They used concepts of multiple testing for controlling the probability of choosing the true model. They referred to the new estimation method as One Covariate at a Time Multiple Testing OCMT.

Setting the identity of covariates $X_j = X_1, X_2, \dots, X_p$ to have signal variables and to simplify the procedure and for the sake of denoting the total of covariates in the regression model as $S_p = \{x_{ij}, j = 1, 2, \dots k, \dots . k^*, \dots, p\}$. Here, there are groups of covariates; k represents all covariates hold $(\beta_j \neq 0)$, while the second group $\{k + 1, \dots, k^*\}$ stands for the covariates that hold $(\beta_j = 0)$,but they possess an

impact of regression model through the net impact, while the remaining $p - k^*$ covariates characterize the covariates that have a marginal or net impact on the regression model [5]. The idea is the addition of some covariates from the second group to the first group by a number of testing iterations, which select the most statistically significant regressors. Subsequently, k is bounded but unknown, we set $h = 1, 2, .. p$. Here Pesaran and Smith [17] express the mean net impact as follows.

$$\theta_j = \sum_{h=1}^{p} \beta_h \sigma_{jh} \qquad j=1,2,\dots,p \qquad \dots (1)$$

The meaning of net impact is the impact that may be caused by one covariate to other covariates. As the test was completed on hypothesis $H_0: \beta_j = 0$ Vs $H_0: \beta_j \neq 0$ as we focus on the net impact. Suppose we have n observations on Y and p covariates on X. In the first stage, we can rewrite the linear model as follows [5].

$$Y = X\emptyset + \varepsilon \qquad \dots (2)$$

Where $\emptyset_j = \frac{\theta_j}{\sigma_{jj}}$ and θ_j is defined in (1), here the parameter $\emptyset$ is regarded as the marginal and the net impact; the idea of replacing the coefficient factor β with $\emptyset$ is that the last one is considered the marginal and net impact on the dependent variable Y and might be a better measure that changes from iteration to other because the remaining variable after each selection. We can denote the t-ratio of $\emptyset$ for the model in (2) to test the hypothesis $H_0: \emptyset_j = 0$ Vs $H_1: \emptyset_j \neq 0$ as $t_{\emptyset j}$ in the following form [3].

$$t_{\emptyset j} = \frac{\widehat{\emptyset}_j}{s.e.(\widehat{\emptyset}_j)} \qquad \dots (3)$$

Pesaran and Smith [17] derive an estimation for the parameter $\emptyset$ through using asymptotic properties as follows:

$$\widehat{\emptyset}_j = \gamma_j (X_j' M X_j)^{-1} X_j' M Y \qquad \dots (4)$$

Where $var(\widehat{\emptyset}_j) = \gamma_j (X_j' M X_j)^{-1}$ and γ_j stand for the variance parameter by using net impact formula, then by using the estimation of the parameters which can be written (3) as follows [17].

$$t_{\emptyset j} = \sqrt{\gamma_j (X_j' M X_j)^{-1}}\, X_j' M Y \qquad \dots (5)$$

Where $M = I_n - \tau_n \tau_n'/n$, τ_n is $n \times 1$ ones vector and $\gamma_j = \sum_{h=1}^{p} \widehat{\sigma}_{jh}^2$ and the estimation of the marginal covariance will be :

$$\widehat{\sigma}_{jh}^2 = \frac{R_j' R_h}{n} \qquad \dots (6)$$

Where $R_j = \left[I_n - X_j (X_j' X_j)^{-1} X_j' \right] Y$ and for each covariate tested in (5) we reject $H_0: \emptyset_j = 0$ if $|t_{\emptyset j}| > C_{(p,\alpha)}$ where $C_{(p,\alpha)}$ is the critical value as follow [12].

$$C_{(p,\alpha)} = \Phi^{-1}\left(1 - \frac{p_r}{2cp^{\alpha}}\right) \qquad \dots (7)$$

Where Φ^{-1} is the inverse standard normal distribution, the test has the tendency to be standard normal distribution, which can be attributed to the extreme size of sample understudy function for positive constants c, α. Choosing critical value is essential since it has rule over the power of the selection process [15].

At the end of the first stage of multiple selection procedure, there will be selection of all covariates that hold $\emptyset \neq 0$ and suppose k is the number of the selected covariates in the first stage. Let X_{k1} be the matrix that has all the selected covariates in first stage, while the rest of $p - k1$ covariate will be in the matrix X_R. At the following stage, the model for the rest of the covariates will be set as the same in first stage and there will be rewriting for the mean net impact.

$$\theta_j = \sum_{h=k1+1}^{p} \beta_h \sigma_{jh} \qquad j=1,2,,\dots,p-k \qquad \dots (8)$$

The regression model will be set as the same in the first stage, for the rest covariates [5].

$$Y = X_R \emptyset + \varepsilon \qquad \dots (9)$$

Now, the second iteration can be made of multiple testing for the model in (9) by repeating the t-ratio test in (5) as there will be a change of the mean net impact.

Suppose that X_{k2} be the matrix that has all the selected covariates in the second stage the t-ratio test will be repeated for the other stages until there are no covariates to be selected by the t-ratio test that eliminates the covariates that have no marginal or net impact on the dependent variable. Then, we set X_k to be the matrix that contains the total number of covariates that selected in all stages of multiple testing procedure

$$X_k = [X_{k1}, X_{k2}, X_{k3}, \dots.]$$

In the last step, the OCMT estimator for the regression model that contains all the selected covariates will be the ordinary least squares estimator OLS as follows.

$$\widehat{\beta}^{OLS} = (X_k' X_k)^{-1} X_k' Y \qquad \dots. (10)$$

Finally, the OCMT estimator will be [5].

$$\widehat{\beta}^{OCMT} = \begin{cases} \widehat{\beta}^{OLS} & (\beta \neq 0) \\ 0 & \text{Otherwise} \end{cases} \qquad \dots (11)$$

4- Proposed Estimator

The multiple test procedure is influential and used to deal with high dimensional data and Big Data sets; the OCMT estimator offers a fast and easy method to apply multiple test procedure via the mean net impact and the t-ratio for the selection of the covariates that possess a marginal and net impact on Y [22]. Here, we suggest a similar approach to the OCMT procedure by using a test of ratio of quadratic forms in normal variables as a selection test for the statistically significant covariate. The test was submitted for the first time by Geoman et al [9], who introduced a score test depending on the ratio of quadratic forms, and this test is not degenerate under the high dimension conditions, so it is appropriate to be used with Big Data sets [22]. Firstly, we make a definition of the ratio of the quadratic form test that will be used subsequently in the multiple testing procedure for selecting the covariates that have an influence on the dependent variable. Suppose we have linear regression model, and we desire to test the hypothesis $H_0: B_j = 0$ Vs $H_1: B_j \neq 0$, Geoman derives a test statistic under high dimension conditions beginning with the following quadratic form test [7].

$$Q = n^{-1}(Y - X\beta)' ZZ' (Y - X\beta) \qquad \dots. (12)$$

Where Z here is the standardized matrix of X and the division by n is beneficial for avoiding degeneracy as the large sample size, [11] Q is quadratic form that follows Chi-square distribution $\chi^2_{(r)}$ with r degree of freedom and to surpass the power of testing for the test statistic in (12). The test on the nuisance parameter is divided as follows: [8]

$$Q = \frac{(Y-X\beta)'ZZ'(Y-X\beta)}{n\sigma^2} \quad \dots (13)$$

Under the null hypothesis and the large size of the sample as $n \to \infty$, Geoman [7] advocated using a pivot approximation for the test statistic in (13) to be suitable in the linear model case as following.

$$S = \frac{Q}{E(Q)} = \frac{(Y-X\beta)'ZZ'(Y-X\beta)}{\text{trace}\,(Z'WZ)} \quad \dots (14)$$

Where $W = \sigma^2 I$ and the test statistic in (14) rely on obtaining σ^2 and this could create some problems of singularity for the matrix $X'X$ as there are Big Data sets under study, so the denominator of the test statistic in (14) can be expressed as follows [19].

$$\text{trace}\,(Z'WZ) = (Y - X\beta)'D(Y - X\beta) \quad \dots (15)$$

By substituting (14) in (13) we can get.

$$S = \frac{(Y-X\beta)'ZZ'(Y-X\beta)}{(Y-X\beta)'D(Y-X\beta)} \quad .. (16)$$

Where D is the diagonal matrix of $Z'Z$; the test statistic in (16) is suitable for the Big Data condition because we avoid obtaining σ^2. Lastly, by substituting the model parameters with the estimated ones, the test statistic will be as follows [18].

$$S = \frac{(Y-X\hat{\beta})'ZZ'(Y-X\hat{\beta})}{(Y-X\hat{\beta})'D(Y-X\hat{\beta})} \quad \dots (17)$$

The expression on (17) will be used in the multiple test procedure for selecting covariates that can impact over Y; the test statistic in (17) has F distribution since it is a ratio between two quadratic forms with a certain degree of freedom and a significant level [9], and we reject $H_0: B_j = 0$ if $S \notin \left[F_{(r_1,r_2,\alpha)}, F_{(r_1,r_2,1-\alpha)}\right]$. The concept behind our suggested estimator is to employ a univariate version of the test statistic in (17) to examine each covariate's marginal and net impact. Now by recalling the OCMT procedure in the first stage, we set the model as follows [19].

$$Y = X\psi + \varepsilon \quad .. (18)$$

Where $\psi = \frac{\theta_j}{\sigma_{jj}}$ and θ_j as in (12) by exchanging the linear model parameter with ψ, we can test the marginal and net impact of each covariate as the new parameter consider them on the model (18) 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 [21].

$$S_j = \frac{(Y-X_j\hat{\psi}_j)'Z_jZ_j'(Y-X_j\hat{\psi}_j)}{(Y-X_j\hat{\psi}_j)'D(Y-X_j\hat{\psi}_j)} \quad \dots (19)$$

Additionally, by recalling the estimation of the regression parameter in (4) [17], the following can be written.

$$\hat{\psi}_j = \gamma_j\left(X_j'MX_j\right)^{-1}X_j'MY \quad \dots (20)$$

Where $M = I_n - \tau_n \tau_n' / n$, $\gamma_j = \sum_{h=1}^{p} \hat{\sigma}_{jh}^2$, $\hat{\sigma}_{jh}^2 = \frac{R_j' R_h}{n}$, and

$R_j = \left[I_n - X_j (X_j' X_j)^{-1} X_j'\right] Y$. By the end of the first stage, let us suppose that X_{m1} be the matrix that holds all the covariates that is statistically important and X_l represents the matrix that contains the other $p - m1$ that were not important and out of selection in the first stage. Now, we set the model and the mean net impact [5].

$$\theta_j = \sum_{h=m1+1}^{p} \beta_h \sigma_{jh} \qquad j=1,2,\dots,p-m1 \qquad \dots (21)$$

$$Y = X_l \psi + \varepsilon \qquad \dots (22)$$

The procedure will be repeated through the application of the ratio test as in (19) until we have no statistically important covariates to be added. Let us set $X_m = [X_{m1}, X_{m2}, \dots]$ to be the matrix that contains all the significant covariates at the end of the last stage then our proposed estimator will be the ordinary least square estimator.

$$\hat{\beta}^{PE} = \begin{cases} \hat{\beta}^{OLS} & (\beta \neq 0) \\ 0 & \text{Otherwise} \end{cases} \qquad \dots (23)$$

5- Criteria of Comparison

There are a variety of statistical comparison procedures that are based on a certain assumption or theoretical foundation [1]. We have chosen the Euclidian Distance, which can be a way of comparison because of the high dimensions of data and the different types of data under study. Euclidian Distance is a useful method for comparing the vectors of estimators that does not require any theoretical background and has a formula for a $(p \times 1)$ estimator vector as follows.

$$ED(\beta) = \frac{\sqrt{\beta_1^2 + \beta_2^2 + \cdots + \beta_p^2}}{p} \qquad \dots (24)$$

6- Results and Discussion

We used a simulation study in our dissertation to analyze the performance of the OCMT and our proposed estimator PR. Simulation studies are highly important to support the work and recommendations of researchers. There were 400 variables generated with a sample size equal to 20000, the variables are standard normal distributed with zero mean and variance equal to one, so here our X matrix will be 20000×400, $p = 400,, n = 20000$ to be close to be Big Data set of variables, and the generating process will be repeated 300 times. OCMT will be calculated from equation (11), our proposed estimator PR will be attained from equation (23), a comparison made by using Euclidian distance ED as in equation (24).

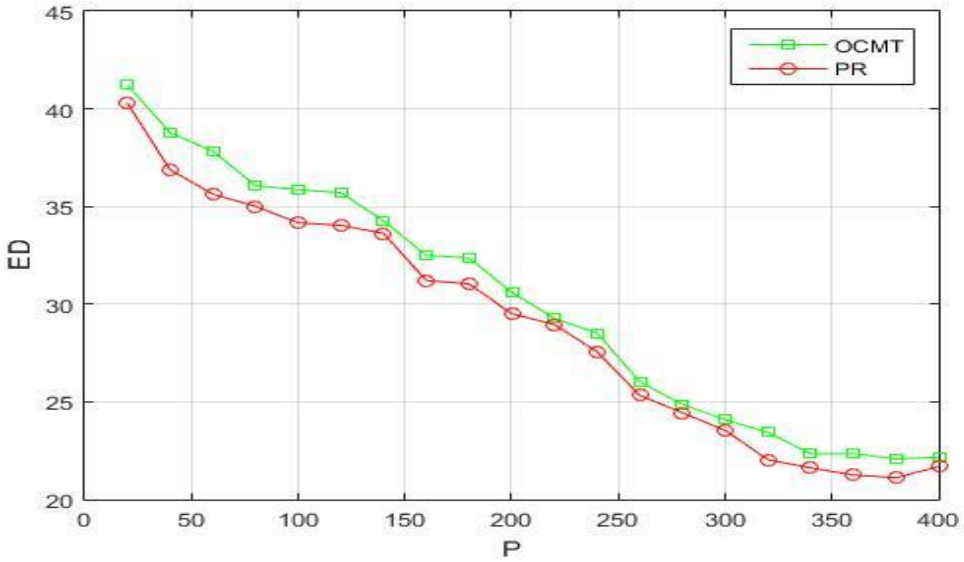

Figure (1) Euclidian Distance (OCMT, PR)

From Figure (1), it can be seen that the OCMT and PR estimators are doing well with a good preference for PR estimator. It is similar when the p gets larger p = 200 OCMT estimator doing very well and PR estimator is the best , when p becomes extremelylarge p = 400 we realize that proposed estimator PR is the best estimator.

Table (1) Euclidian Distance (ED) for the Estimation Methods (OCMT, PR)

P	OCMT	PR	P	OCMT	PR	P	OCMT	PR
20	41.220	40.295	160	32.489	31.208	300	24.085	23.539
40	38.781	36.872	180	32.367	31.035	320	23.431	22.034
60	37.823	35.633	200	30.608	29.512	340	22.341	21.616
80	36.072	35.009	220	29.263	28.952	360	22.345	21.253
100	35.865	34.163	240	28.518	27.560	380	22.075	21.114
120	35.700	34.034	260	26.037	25.327	400	22.165	21.683
140	34.263	33.618	280	24.860	24.441			

Here, tables will be explained in the simulation study in accordance with the number of variables p as follows: When p = 20 – 100 and from Table (1) , there is good performance for PR estimator and it is almost the same for OCMT estimator, when the number of variables p =180 – 280 when there is still the fine performance of the proposed estimator PR and OCMT estimator. Moreover, when p = 300-400, PR estimator still performs satisfyingly and it is the same for OCMT estimator; from the result of the simulation study, it's clear that PR is the best estimator. Before discussing the data details, a short explanation of the concept of child labor needs to be introduced and it was calculated for a group of families as follow:

$$Y = \frac{q}{t} \qquad \dots (3.1)$$

Where q is the number of children under 14, who work, and t represents the total number of children. We have acquired a large number of sets of surveys data from the Central Statistical Organization IRAQ to represent 10000 group of families from variety parts of the country, and we calculate the MPI vector (10000×1); we have 340 variables from various types quantitative, ordinal, nominal, etc…

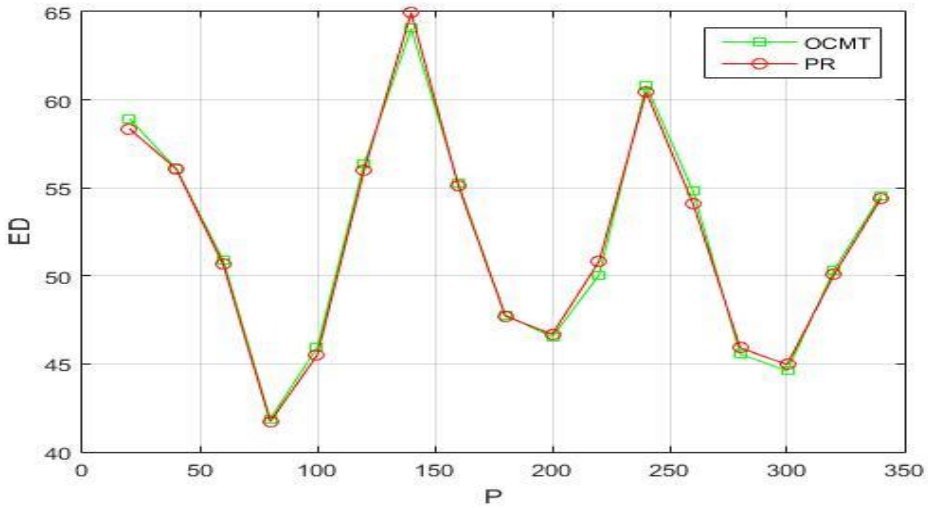

Figure (2) Euclidian Distance (OCMT, PR)

Figure (2) displays OCMT and PR estimators perform fully good, when p = 75, besides the best estimator is PR with a pleasant performance and OCMT estimator performs an agreeable routine, when p = 300, the proposed estimator PR retains to be the best estimator.

Table (2) Euclidian Distance ED for the Estimation Methods (OCMT, PR)

P	OCMT	PR	P	OCMT	PR	P	OCMT	PR
20	58.93351	58.3624	140	64.05336	64.95065	260	54.80652	54.09717
40	56.07598	56.06633	160	55.26968	55.14106	280	45.56152	45.91072
60	50.88382	50.65052	180	47.78088	47.72262	300	44.60261	44.95859
80	41.87114	41.71824	200	46.52746	46.67422	320	50.34595	50.07329
100	45.92844	45.49849	220	50.05477	50.82404	340	54.57629	54.40262
120	56.36501	56.01582	240	60.81665	60.44945			

By adding 10 real variables at a time to calculate the ED for all the five estimators, we will clarify all tables by using real data according to the number of variables p as follows. When p = 20 – 100 and from Table (2), the performance of PR estimator is good and also for the OCMT estimator, when the number of variables p = 100 – 200 PR and OCMT estimators perform pleasantly fine, and when p = 300-340, the great performance of PR and OCMT estimators continues, and PR is the best estimator. As a summary of the result of the real data study, PR is the best estimator for estimating the coefficients of the linear regression model.

7- Conclusions

The findings of the simulation study demonstrate that the suggested estimator PR has very good performance and is very near to the OCMT method when the number of variables is quite large. When real data is used, the performance of the proposed estimator PR and OCMT estimators is the best when the number of variables is quite small, and with Big Data properties it becomes very obvious that the proposed estimator PR and OCMT methods that are dependent on the multiple testing procedure to select the statistically significant variable are the best with little preference for proposed estimator PR.

We recommend our proposed estimator for estimating the coefficient of a linear regression model under Big Data conditions because of its good performance in simulation studies, where the proposed estimator PR gets the smallest values of Euclidean Distance ED when p is small or large; moreover, the proposed estimator PR estimator shows a significant performance when real data were used.

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]. Boyd. D, Crawford. K, "CRITICAL QUESTIONS FOR BIG DATA Provocations for a cultural, technological, and scholarly phenomenon" Information, Communication & Society, Vol.15, No.5, pp. 662-679, 2012.

[3]. Buhlmann. P, Van De Geer. S, " Statistics for High-Dimensional Data: Methods, Theory and Applications" Springer, 2011.

[4]. Chang. W, Grady. N, "NIST Big Data Interoperability Framework: Volume 1, Definitions" Special Publication (NIST SP) - 1500-1 Version 2, 2012.

[5]. Chudik. A, Kapetanios. G, Pesaran. M, "One-Covariate at Time, Multiple Testing Approach to variable selection in High-Dimensional Regression Models" Econometrica, Vol. 86, Issue. 4, pp. 1479-1512, 2018.

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

[7]. Geoman. J, Van De Geer. S, "A Global Test for Groups of Gens: Testing Association With A Clinical Outcome" Bioinformatics, Vol. 20, pp. 93-99, 2004.

[8]. Geoman. J, Van De Geer. S, Van Houwelington. H, "Testing Against a High-Dimensional Alternative" Journal of Royal Statistics Society, Vol.68, pp. 477-493, 2006.

[9]. Geoman. J, Van Houwelington. H, Finos. L, "Testing Against a High-dimensional Alternative in the Generalized Linear Model: Asymptotic Type 1 Error Control" Biometrika, Vol. 98, No. 2, pp. 381-390, 2011.

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

[11]. James. W, Stein. C, "Estimation with Quadratic loss" Proceedings 4th Berkeley Symposium, pp. 361-379, 1961.

[12]. Khorshed. E, Abood, S, "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 Science, Vol. 24, No. 109, pp. 515-534, 2018.

[13]. Kumar. R, Moseley. B, Vassilvitskii. S, Vattani. A, "Fast Greedy Algorithms in Map Reduce and Streaming" ACM Transactions on Parallel Computing, Vol. 2, No. 3, pp. 154-170, 2011.

[14]. Mauro. A, Greco. M, Grimaldi. M," What is Big Data? Consensual Definition and Review of Key Research Topics" International Conference on Integrated Information.AIP Conference Proceedings. Proc. 1644, pp 97-104, 2015.

[15]. 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.

[16]. Nikolova. M, "Local Strong Homogeneity of Regularized Estimator" SIAM Journal of Applied Mathematics, Vol. 61, No. 1, pp. 633-658, 2000.

[17]. Pesaran. M, Smith. R, "Signs of Impact Effects in Time Series Regression Models" Economics Letters, Vol. 122, pp. 150-153, 2014.

[18]. 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.

[19]. Salih. A, Hmood. M, "Big Data Analysis by Using One Covariate at a Time Multiple Testing (OCMT) Method: Early School Dropout in Iraq" Int. J. Nonlinear Appl. Issue. 12, No. 2, pp. 931-938, 2021.

[20]. Schroeck, M, Shockley. R, Smart. J, Romero-Morales. D, Tufano. P, "Analytics: The Real-World Use of Big Data" IBM Global Services Route 100 Somers, NY 10589 U.S.A, pp. 1-19, 2012.

[21]. Van Der Vaart. A, "Asymptotic Statistics" Cambridge: Cambridge Series in Statistical and Probabilistic Mathematics, 1998.

[22]. Xu. K, "A new nonparametric test for high-dimensional regression Coefficients' Journal of Statistical Computation and Simulation, Vol. 5, pp. 855-867, 2017.

Appendix 1 :- Variables under Study

No	Variable	NO	Variable
1	Household number	41	Region
2	persons number	42	Mother's number up 48
3	Age Average	42	Father's number up 48
4	Sex Ratio	44	Education of household head
5	Men number	45	Functional difficulties
6	number of woman age 15 - 49	46	Health insurance
7	number of man age 15 - 49	47	Age at beginning of school year
8	number for children age 0-4	48	Mother’s education
9	Member age 0-17	49	Mother's disabilities (age 18-49 years)
10	Is natural mother alive	50	Father’s education
11	Does natural mother live in H	51	Household sample weight
12	Natural mother's number in H	52	Combined wealth score
13	number natural mother live O	53	Wealth Quintile
14	Is natural father alive	54	Percentile Group of com1
15	Does natural father live in H	55	Wealth Quintile Urban
16	Natural father's l number in H	56	Percentile Group of urb1
17	number natural father live O	57	Rural wealth score
18	number caretaker for children 0-17 age	58	Wealth Quintile Rural
19	Mothers age average	59	Percentile Group of rur1
20	fathers age average	60	Primary sampling unit
21	Age 4 and above	61	Stratum
22	Early Childhood Education	62	Household ID
23	Highest level of education attended	63	Individual ID
24	Highest grade attended at that level	64	Highest educational level attended
25	Highest grade completed at that level	65	Highest year of education completed
26	Age 4-24	66	number of years of education
27	Early Childhood Education	67	Child education u 6
28	Attended school during last 5y	68	years of education u 6
29	Level of education attended	69	at least one member with 6 years of edu
30	Grade of education attended	70	Attended school during current year
31	Attended public school	71	child schooled
32	School tuition in the current school year	72	missing school attendance for at least 2/3
33	Material support in the current school year	73	Household has children in school age
34	Attended school	74	child no attending private
35	Level of education attended private	75	child not attended private
36	Grade of education attended previous	76	all school age children to class 8 in school
37	Child under 10 N A	77	Woman's number NA
38	Child under 17 NA	78	Women BH
39	Area of household	79	Total child death for each women
40	Region/Governorate	80	Total child death last 5 years

استخدام اختبار النسبة للصيغة التربيعية لتقدير انموذج الانحدار الخطي للبيانات الكبيرة مع تطبيق عملي

أ.م. أحمد مهدي صالح
قسم الإحصاء / كلية الإدارة والاقتصاد
جامعة واسط

أ.د. مناف يوسف حمود
قسم الإحصاء / كلية الإدارة والاقتصاد
جامعة بغداد

Received:18/8/2021 Accepted: 22/8/2021 Published: March / 2022

مستخلص البحث:

في هذا البحث تم تقديم مقدر مقترح لتقدير معالم انموذج الانحدار الخطي في ظل وجود البيانات الكبيرة. حيث تنوع المتغيرات في البيانات الكبيرة تجلب العديد من التحديات التي تغري الباحثين لإيجاد طرق تقدير جديدة لمعالم انموذج الانحدار الخطي. تم جمع بيانات من الجهاز المركزي للإحصاء. وتم اختيار البيانات المتعلقة بعمالة الأطفال. حيث تعد عمالة الأطفال من الظواهر المهمة التي يعاني منها المجتمع والتعليم وتوثر في مستقبل الأجيال القادمة. تم اختيار طريقتين لتقدير معالم انموذج الانحدار في البيانات الكبيرة وهي طريقة OCMT بالإضافة الى الطريقة المقترحة. وتم استعمال المسافة الاقليدية كمعيار للمقارنة بين الطريقتين . وكان الهدف هو بيان الطرق الأفضل لتقدير معالم انموذج الانحدار في البيانات الكبيرة. وتم التوصل الى افضلية الطريقة المقترحة لتقدير انموذج الانحدار في البيانات الكبيرة. وان البيانات تمثل مؤشرات حيوية لمجاميع من العوائل من مختلف محافظات العراق. وتم التركيز على المتغيرات التي تدفع الى تنامي ظاهرة عمالة الأطفال الاقتصادية منها والصحية.

الآثار الاجتماعية: التركيز على بعض المتغيرات الاجتماعية مثل وجود الاب او امراض التوحد التاي يعاني منها الأطفال والتي تدفعهم في بعض الأحيان الى ترك الدراسة والبحث عن عمل .

نوع البحث: ورقة بحثية.

المصطلحات الرئيسة للبحث: البيانات الكبيرة . عمالة الأطفال . الفقر متعدد الابعاد . انموذج الانحدار الخطي. OCMT .

*البحث مستل من اطروحة دكتوراه