TEST Engineering & Management

November-December 2019
ISSN: 0193-4120 Page No. 783 - 788

Measuring Performance of Distance-based Regression for Skewed Data

Nor Hisham Haron[1], Nor Aishah Ahad[2], Nor Idayu Mahat[3]

[1]Lecturer at Mathematics and Statistics Department, School of Quantitative Sciences, Universiti Utara Malaysia
[2]Senior lecturer in School of Quantitative Sciences, Universiti Utara Malaysia
[3]Senior lecturer in School of Quantitative Sciences, Universiti Utara Malaysia

Article Info
Volume 81
Page Number: 783 - 788
Publication Issue:
September-October 2019-10-21

Article History
Article Received: *3 January 2019*
Revised: *25 March 2019*
Accepted: *28 July 2019*
Publication: *25 November 2019*

Abstract: Cuadras introduced a Distance-based regression (DBR) in 1990as an unbiased regression model that suitable to use in mixed-type of independent variables. DBR is similar to classical linear regression (CLR), but it utilizes distance measures as independent variables instead of raw values. Earlier study on DBR has limited the focus on understanding the performance of DBR when the data are normally distributed, hence it performances in skewed data remain questionable. This study attempts to answer such question by comparing the performance of DBR with bootstrap linear regression (BLR), in simulated data sets, which contain either continuous independent variables or mixed type of independent variables where residuals were set to follow gamma distribution. The simulations consider the number of sample size, n and number of independent variables, p. Small (n = 10), medium (n = 40) and large (n = 100) with p = 2 and p = 3.The investigation was set up in a simulation study, aiming to compare the performance of DBR over BLR based on the value of adjusted R-square (adjR2), Bayesian information criterion (BIC) and power. Power is the percentage of p-value for the model that less than 5% significance level. The main objective for this study is to see in what circumstances DBR is suitable to use. The findings indicate that DBR performed better thanBLR in all cases of numerical independent variables and mixed-type of independent variables. We also found that DBR performed better across all tested sizes of sample.

Keywords: *Bootstrap Linear Regression, Distance-based Regression, Gamma Distribution, Gower Distance,Skewed Data.*

I. INTRODUCTION

Regression analysis is a statistical tool that has commonly used to indicate relationship between dependent and one or more independent variable. Beside, this tool has been often used for data description, parameter prediction and control. For data description, most engineers and statisticians use regression model to describe patterns of data in hand, which often presented in term of mathematical. In case of estimating an unknown parameter, regression analysis becomes as one of possible tools especially when the parameter of interest has shown some explainable pattern like a

November-December 2019
ISSN: 0193-4120 Page No. 783 - 788

linear trend. Since regression is a mathematicalmodel, hence it offers one to predict value of a dependent variable given information of independent variables. Finally, regression model can be used in controlling the dependent variable by chose the suitable values in independent variables. Despite of these There are four assumptions on the error term i.e. normality, homoscedasticity, independency and linearity [1]. Violation of one of these assumptions may make the regression model inefficient, biased or misleading then it will make the prediction of new observations deviate from exact value. The error term can be not normal if one or more extremes value occur in the data set. The non-parametric method is one of alternatives to solve problem in regression for non-normal data. The most popular techniques in non-parametric regression are local linear, nadaraya-watson kernel and bootstrapping regression (BLR).Then, this study wants to determine whether DBR is suitable for non-normal data. DBR is an approach based on distance and a tool that can be applied to categorical or continuous or mixed explanatory variables. [2] fundamentally defined a model of DBR for a linear regression model. DBR also can be used as a method to estimate the parameters.The extended study to determine an appropriate distance based on the type of data found that Gower distance is suitable [3].

Basically DBR is same as CLR, but the value of $\boldsymbol{X_{(k)}}$ is obtainedusing metric scaling on $n \times n$ distance matrix. In general, the steps in model building in DBR consisting of two major steps: (i) from a raw data of explanatory variables, we calculate the distance between observations using the suitable distance function and (ii) then we obtain the matrix $\boldsymbol{X_{(k)}}$ of principle coordinates. Since we have the new explanatory variables, $\boldsymbol{X^*}$ (in distance form), we perform an ordinary least squares regression of dependent variable. The $\boldsymbol{X^*}$is a column vector of eigen vector of $\boldsymbol{X_{(k)}}$. [1] used Gower's distance to estimate the parameter of regression line and measure the value of R^2 and cross validation (CV). This study tries to investigate the performance of DBR for data that have non-normal error distribution.Despite of R2 and CV, we measured performance of DBR base on the values of adjusted $adjR^2$, BIC and power. $AdjR^2$ is a tool to measure the percentage of the variation of explanatory variables describe the dependent variable. It is similar with R^2but the value will increase if added independent variable is significant to the model. The BIC is tool to select the best method among several models. The best model gives the smallest value of BIC. Whereas the power is the percentage of *p*-value that fit to the model constructed. In this study, the significance level was set at 5% as the limit of *p*-value. Any *p*-value lower than 5% is significance and vice versa. In the process of investigation, the non-normal error, $\boldsymbol{e_i}$ was simulated followed a gamma distribution; $e_i \sim \Gamma(\alpha, \beta)$. The dependent variable, $\boldsymbol{Y}$ was generated based on uniform distribution; $Y \sim U(a, b)$ and different type of explanatory variables (i.e.: categorical and continuous). The explanatory variables were either all continuous or mixtures of continuous and categorical variables. The categorical variables are in the form of binomial, nominal or ordinal.

II. METHODOLOGY

The aim of this study is to investigate the performance of DBR by comparing the value of adjR2, BIC and power with BLR.In this study, the explanatory variables are categorized into two groups, i.e. all continuous and mixed variables.Then for each category, the number of independentvariables, *p* is varies from two to three ($p = 2$ and $p = 3$) and only one dependent variable. We simulate the data in 1000 repetition [4] using R program based on different sample size; $n = 10$ (small), $n = 40$ (medium) and $n = 100$ (large). Figure 1 shows the study framework of our study.

November-December 2019
ISSN: 0193-4120 Page No. 783 - 788

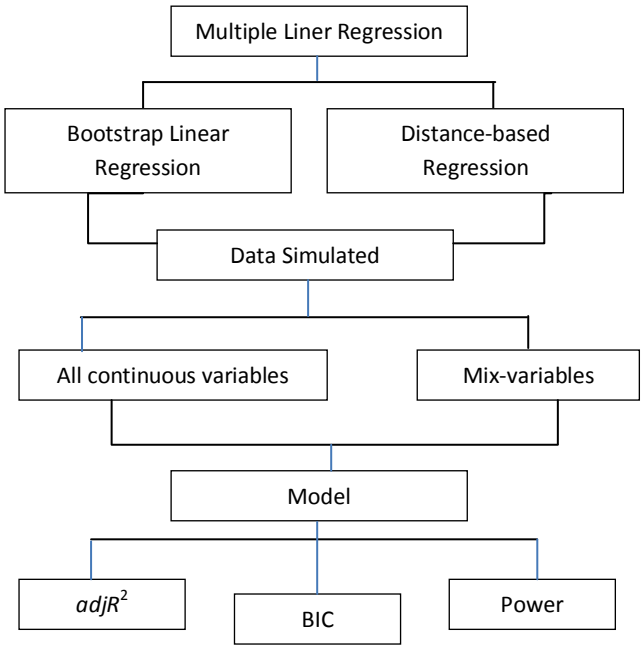

Fig 1: Experimental Framework of the Study

Data Generation

As discuss in previous section, one of the assumption for regression analysis is normality in the error term.Gamma distribution was generated to fulfill the condition of non-normal data. The gamma distribution is depend on the shape and the rate. In this study, we set the gamma distribution with shape = 1 and rate = 1.

Algorithm of DBR

In constructing the DBR model, four major phases were involved. First phase is transformation the data into distance matrix of raw data. Suppose that we have a multiple linear regression of $\hat{\boldsymbol{y}} = \boldsymbol{b_o} + \boldsymbol{b_i w_i}$ .From simulated variable, $\boldsymbol{y}_i$ and $\boldsymbol{w}_i$, then calculate the distance matrix for $\boldsymbol{w}_i$, $\boldsymbol{Dg}$and$\boldsymbol{Dg}$ is based on the type of data. We used gower distance as suggested by Cuadras. For the second phase, we find the matrix $\boldsymbol{X}$, where

$\boldsymbol{B} = \boldsymbol{HAH}$

$\boldsymbol{A} = -\frac{\boldsymbol{Dg}}{\boldsymbol{2}}$,

$\boldsymbol{H}$ is the hat matrix

Assume that $\text{rank}(\boldsymbol{B}) = m$, then

$\boldsymbol{XX}' = \boldsymbol{B}$

$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{V} = \text{diag}(\lambda_1, \cdots, \lambda_m)$, $\boldsymbol{X}$ is an$\times m$ matrix of rank m

$\lambda_i =$ positive eigenvalue of$\boldsymbol{B}$

In this phase, the transformation of the raw data, $\boldsymbol{w_i}$ is based on the gower distance. Next phase is selecting the $\boldsymbol{X}$*from $\boldsymbol{X_{(k)}}$, where $\boldsymbol{X}$* is the eigenvectors of $\boldsymbol{B}$ respecting to the positive eigenvalues of $\boldsymbol{B}$.

The final stage is constructing the regression model in form of distance-based, $\hat{\boldsymbol{y}} = \hat{\boldsymbol{\beta}}_0 + \sum_{i=1}^{n} \hat{\boldsymbol{\beta}}_1 \boldsymbol{X_i^*}$

III. RESULT AND DISCUSSION

The results of analysis based on simulated data are displayed in table 1 until table 3.

Table1: Result of average $adjR^2$, BIC and Power for All Continuous Independent Variables, p

Sample size, n	p = 2		p = 3	
	BLR	DBR	BLR	DBR
$adjR^2$				
10	76.53	79.16	53.03	91.46
40	73.02	72.79	45.81	71.59
100	14.04	17.37	33.60	50.49
BIC				
10	40.17	40.22	85.83	39.80
40	253.37	253.37	298.33	256.39
100	682.55	682.53	721.92	685.26
Power				
10	82.7	85.9	61.3	94.1
40	98.8	98.3	85.0	97.0
100	75.2	80.6	88.7	97.3

For all continuous independent variables, the values of $adjR^2$ for both cases p = 2 and p = 3 shows that DBR gives the better result as compared to BLR. In small sample size, n = 10 and p = 2, BLR shows 79.53% of the variation in independent variables describes the dependent variable. Whereas DBR show 79.16% of independent variables describe dependent variable. For n = 10 and p = 3, $adjR^2$ for BLR resulted 53.03% and DBR 91.46%. When we

TEST Engineering & Management

November-December 2019
ISSN: 0193-4120 Page No. 783 - 788

increase the n to medium and large (n = 40 & 100), we observed the same pattern showed where DBR performed better than BLR.

For the value of BIC, the smallest value of BIC shows the data fit the model. For all n and p = 2 BLR gives the better result compared to DBR. But when p = 3, DBR gives the better result for all n. In terms of power, all conditions shows that DBR is better than BLR.

Table 2: Results of average $adjR^2$, BIC and Power for

n		10	40	100
Case1: one continuous and & binomial				
$adjR^2$	BLR	85.18	60.44	38.15
	DBR	86.12	61.97	39.23
BIC	BLR	41.85	257.19	683.26
	DBR	41.74	257.16	683.25
Power	BLR	90.9	91.9	89.3
	DBR	91.3	93.8	92.4
Case2: one continuous & one nominal				
adj^R2	BLR	90.60	67.40	47.27
	DBR	87.83	60.32	39.48
BIC	BLR	40.04	255.58	686.82
	DBR	42.26	256.67	687.97
Power	BLR	94.1	94.3	95.1
	DBR	93.8	90.2	91.4
Case3: one continuous & one ordinal				
adj^R2	BLR	89.35	69.82	48.13
	DBR	86.18	62.19	39.55
BIC	BLR	40.26	256.07	692.16
	DBR	41.88	257.19	683.25
Power	BLR	92.2	95.3	95.2
	DBR	91.9	92.7	90.3

The result for p = 2 with one continuous and one binomial shows that DBR performed better in the value of $adjR^2$. For n = 10 shows 86.18% the variation in independent variables describe the dependent variable whereas BLR gives the value of 85.18%. for n = 40, DBR shows the value of 61.97% while the BLR give 60.44%. for n = 100, DBR shows the value of 39.23% while BLR gives the value of 38.23%. In term of BIC and power, DBR and BLR give the compatible results for each sample size.

A different result shows were obtained for the second case with one continuous and one nominal variables. In this case, BLR shows better performance compared to DBR. The value of $adjR^2$ of BLR shows the percentage of 90.60%, 67.40% and 47.27% for n = 10, 40 and 100 respectively. DBR only give 87.83%, 60.32% and 39.48%. In term of BIC, BLR shows compatible results but always quite lower than DBR. The same results can be observed for the third case that contained one continuous and one ordinalvariables.

Table 3: Results of average $adjR^2$, BIC and Power for case of p = 3

n		10	40	100
Case4: one continuous &two binomial				
$adjR^2$	BLR	90.63	72.71	48.32
	DBR	91.41	74.02	49.30
BIC	BLR	38.92	251.08	687.11
	DBR	38.78	251.06	687.08
Power	BLR	92.9	96.2	95.7
	DBR	93.9	98.3	96.0
Case5: one continuous &two ordinal				
adj^R2	BLR	36.40	46.10	16.33
	DBR	52.23	18.26	15.65
BIC	BLR	85.32	376.86	949.25
	DBR	94.38	381.62	953.78
Power	BLR	16.8	94.3	86.2
	DBR	52.7	58.6	80.8
Case6: two continuous & one binomial				
adj^R2	BLR	89.75	71.00	49.33
	DBR	91.34	72.23	50.62
BIC	BLR	40.07	256.44	685.26
	DBR	39.97	256.42	685.23
Power	BLR	92.4	96.1	96.8
	DBR	94.5	97.0	96.7

The result for p = 3are shown in table 3. In case of one continuous and two binomial variables, DBR shows better performance as compared to BLR in all values of $adjR^2$, BIC and power. But for the

November-December 2019
ISSN: 0193-4120 Page No. 783 - 788

case of one continuous and two ordinal variables, DBR performed better only when the *n* is small (*n* = 10) in the value of $adjR^2$. DBR gives the result of 52.23% as compared to BLR only36.40%. But, the BIC for BLR was performed better than DBR. It shows the value of 94.38 compared to BLR of 85.32. When the sample size increase to medium and large size, BLR give the better results for all measurement tools. For case two continuous and one binomial variables, DBR shows better result for each values of adjR2, BIC and power for all sample sizes.

To validate the models of the simulated study, a real data were tested. We used three sets of real data: (i) two independent variables that consist of all continuous variable; (ii) two independent variables that consist of one continuous and one binomial variable; and (iii) three independent variables, which is, consist of one continuous, one binomial and one ordinal. The result of these three real data shown in Table 4.

Table 4: Result of $adjR^2$, BIC and Power for Real data

Data 1: All continuous independent variables, *n* = 24		
Method	Measurement	
	$adjR^2$	BIC
BLR	78.43	125.11
DBR	86.34	120.09
Data 2: One continuous & one binomial, *n* = 57		
Method	Measurement	
	adjR^2	BIC
BLR	81.23	235.41
DBR	87.44	230.43
Data 3: one continuous, one binomial & one ordinal, *n* = 120		
Method	Measurement	
	*adjR*2	BIC
BLR	90.04	374.34
DBR	91.23	370.26

Based on table 4, the finding are show that the real data is consistent with the simulated data. In all cases, DBR gives the best result for all level of sample size.

IV. CONCLUSION

Based on this initial study, for the data that consist of all continuous and mixed independent variables, distance-based regression (DBR) out performed boot strapping linear regression (BLR). In conclusion, DBR and BLR are suitable to be use for non-normal data in small, medium and large sample size.

REFERENCES

[1]. D. C. Montgomery and E. A. Peck, *Introduction to Linear Regression Analysis*. New York: John Wiley & Son, 1991, pp. 67-113.

[2]. C.M. Cuadras and C. Arenas. *Communications in Statistics – Theory and Methods*, ***19,*** 2261 – 2279.(1990).

[3]. C.M. Cuadras, C. Arenas and J. Fortiana. *Communications in Statistics – Simulation and Computation*, ***25(3),*** 593 – 609.(1996).

[4]. E. Boj, M.M. Claramunt, J. Fortiana and A. Vidiellaj (2001). The use of distance-based regression and generalized linear models in the rate making process. An empirical study.

[5]. Retrieved from www.imub.ub.es/publications/preprints/pdf/Boclaforvi.pdf

[6]. E. Boj, M. M. Claramunt and J. Fortiana, J. (2006). Bootstrapping pairs in distance-based regression. Retrieved from http://EconPapers.repec.org/RePEc:bar:bedcje:2006154.

[7]. E. Boj, A. Grane, J. Fortiana and M.M Claramunt. *Statistics and Econometrics Series* **14,** working paper, 6-35.. (2006).

November-December 2019
ISSN: 0193-4120 Page No. 783 - 788

[8]. E. Boj, M.M. Claramunt and J. Fortiana. *Communications in Statistics – Simulation and Computation*, ***36*****:** 87 – 98.(2007).

[9]. E. Boj, P. Delicado and J. Fortiana. (2008). Local linear functional regression based on weighted distance-based regression. Retrieved from http://www-eio.upc.es/~delicado/my-public-files/LocLinFunct.pdf

[10]. E. Boj, P. Delicado and J. Fortiana,. *Computional and Data Analysis*, **54,** 429 – 437. (2010)

AUTHORS PROFILE

Nor Hisham Haron was born in 1972 in Pasir Mas Kelantan, Malaysia. He has received his Bachelor of Statistics (with honors) from Universiti Teknologi Mara, Shah Alam Selangor, Malaysia in 2000. In 2001 he continued his study in Master of Decision Sciences from Universiti Utara Malaysia, Sintok Kedah, Malaysia. At present he is a part time doctorate student from Universiti Utara Malaysia. At the same time, he is a lecturer at Mathematics and Statistics Department, School of Quantitative Sciences, Universiti Utara Malaysia.

Nor Aishah Ahad is a senior lecturer in School of Quantitative Sciences, Universiti Utara Malaysia. She obtained her PhD in Statistics from Universiti Sains Malaysia. Her research interests are in robust statistics, nonparametric methods, statistical modelling and applied statistics.

Nor Idayu Mahat is a senior lecturer in School of Quantitative Sciences, Universiti Utara Malaysia, with major interests in multivariate analysis, statistical modelling, and pattern recognition. She obtained her PhD in Mathematics from University of Exeter, United Kingdom. At present, she is attached as Director at Centre for Testing, Measurement, and Appraisal at the same university. She practices her quantitative discipline in various areas including engineering, business and management, and talent development especially in handling computation with Big Data