MJFAS MALAYSIAN JOURNAL OF FUNDAMENTAL AND APPLIED SCIENCES
PRINT ISSN: 2289-5981 | ONLINE ISSN: 2289-599X

RESEARCH ARTICLE

Predicting the performance of the players in NBA Players by divided regression analysis

Yann Ling Goh [a,*], Yeh Huann Goh [b], Raymond Ling Leh Bin [c], Weng Hoong Chee [a]

[a] *Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Jalan Sungai Long, Cheras, 43000 Kajang, Selangor, Malaysia*
[b] *Department of Mechanical Engineering, Faculty of Engineering, Kolej Universiti Tunku Abdul Rahman, Jalan Genting Kelang, Setapak, 53300 Kuala Lumpur, Malaysia*
[c] *Department of Accountancy, Faculty of Accountancy and Management, Universiti Tunku Abdul Rahman, Jalan Sungai Long, Cheras, 43000 Kajang, Selangor, Malaysia*

* *Corresponding author: gohyl@utar.edu.my*

Article history
Received 18 November 2017
Revised 5 July 2018
Accepted 12 August 2018
Published Online 25 June 2019

Abstract

A divided regression model is built to predict the performance of the players in the National Basketball Association (NBA) from year 1997 until year 2017. The whole data set is divided into five groups of sub data sets and multiple linear regression model is employed to model each of the sub data set. In addition, the relationships among independent variables are checked by using variance inflation factor (VIF) to identify the risk of having multicollinearity in the data. Moreover, non-linearity of regression model, non-constancy of error variance and non-normality of error terms are investigated by plotting residual plots and quantile-quantile plots. Finally, a divided regression model is built by combining the results obtained from the sub data sets and the performance of the divided regression model is verified.

Keywords: Divided regression, multiple linear regression, variance inflation factor

© 2019 Penerbit UTM Press. All rights reserved

INTRODUCTION

In statistical modelling, regression modelling is a statistical tool and process for estimating and modelling the relationship between a response variable (target) and one or more explanatory variables (predictor) [1]. In real life, regression modelling is widely used in the cases for prediction and forecasting which has substantial overlap with the field of machine learning. Linear regression is one of the strongest tools available in statistics and machine learning for predicting output variable (y) given some input variables (x). For instance, the relationship between sales and advertising cost is best studied through regression. Furthermore, regression analysis is used for making comparison among variables measured on different scales that helps data analysts to evaluate the optimal set of variables to be used in building the regression model. In other words, regression analysis provides us the understanding on the way of variation in the response variable when any one of the predictor variables is changed.

There are various types of regression techniques such as linear regression, logistic regression, polynomial regression, stepwise regression, ridge regression, lasso regression and ElasticNet regression [2-4]. In this paper, we are dealing with the case in linear regression. Linear regression is divided into simple linear regression and multiple linear regression where the former has only one independent variable while the latter takes in more than one independent variables as inputs. Simple linear regression uses a best-fit straight line, which is known as regression line to explain a relationship between a response variable and an independent variable [5]. On the other hand, the multiple linear regression model describes a plane in three-dimensional space of response variable and several independent variables [6, 7]. The Least Squares Method is applied in order to obtain the regression coefficients for simple linear regression model and multiple linear regression model.

In big data environment, a big data set that is closed to population can be obtained from advanced computer system and thus we are able to carry out statistical analysis over the big data set to estimate and make inference to the population. However, in the era of big data, we may not able to analyze the entire part of big data due to its huge data volume. In addition, the statistical methods also have computing limitation to manipulate and control massive data set [8-11]. Therefore, in order to reduce the computing load, an analytical methodology for big data analysis in linear regression problem is then proposed which is called divided regression analysis.

The proposed method is to split a huge volume of data into many, said n sub data sets. In each of the sub data set, a multiple linear regression model is constructed for estimating the coefficients of regression parameters; therefore n multiple linear regression models will be built. Hence, the values of regression parameters of entire data set can be estimated by taking combine function for merging the results from sub data set 1 to sub data set n [12].

Yang (2015) has applied linear regression analysis on NBA players to predict their regular season results based on the common basketball statistics. There are two major variables which are being used in building the regression model, which includes the win ratio of team and PER where PER is a metric of measurement on player's effectiveness with a single number. Furthermore, the author has

investigated the correlation between the two variables by plotting scatter plot of variables [13].

Aiken *et al*., (2003) have conducted studies on the relationship between a single response variable and several predictor variables by using multiple regression analysis. Multiple regression analysis can be applied to summarize the relationships of a set of independent variables to a criterion at a single point in time. One general form of the regression equation written for an individual case i is as follows:

$$\hat{Y}_i = b_0 + b_1X_{i1} + b_2X_{i2} + \cdots + b_jX_{ij} + \cdots + b_pX_{ip}$$

where $X_{i1}, X_{i2}, \dots, X_{ip}$ are the scores of case i on the $j = 1, 2, \dots, p$ predictors; $b_1, b_2, \dots, b_p$ are partial regression coefficients and b_0 is the regression intercept. Moreover, they illustrate the use of multiple regression to investigate the relationship of qualitative and quantitative predictor variables [14].

Similarly, multiple linear regression model is applied by Fitrianto et al. (2016) to model the mortality rate of children and related factors in Asia [15]. Correlation coefficient is introduced to figure out the relationships between variables and it is computed by dividing the covariance of the variables by the product of their standard deviation,

$$correlation(X_1, X_2) = \rho_{X_1,X_2} = \frac{cov(X_1,X_2)}{\sigma_{X_1}\sigma_{X_2}},$$

Hypothesis test is carried out to check the correlation between two variables,

$$H_0: \rho_{X_1,X_2} = 0$$
$$H_1: \rho_{X_1,X_2} \neq 0$$

If H_0 is rejected, then there is a correlation between X_1 and X_2.

When there are more than one predictor variables in a regression model, a risk of having the problem of multicollinearity always exists. Hall, Fienberg and Nardi (2011) stated that multicollinearity is a statistical phenomenon where two or more predictor variables in a multiple regression model are highly correlated to one another, causing a great instability of regression coefficients and resulting in misleading interpretation in the estimated regression coefficients [16, 17].

To deal with big data analysis, a split and conquer approach had been studied by Chen and Xie (2014). They mentioned that it is not easy to fit extraordinary large data into a single computer due to the computing burden and high cost. By using split and conquer method, a whole data set is split into non-overlapped small sub data set randomly. Then, every sub data set is analyzed separately and the results obtained for each sub set are combined together by a systematic way to give the final overall statistical inference that contains information of the whole data set. They have used some penalized regression approaches, which are often studied in huge volume data analysis to demonstrate their suggested method. As a result, the final value obtained by this approach is asymptotically equivalent to the result of the entire data set, assuming that there is a super computer, which is able to perform the analysis of the entire data set [18].

Likewise, a similar method was also applied by Tang, Zhou and Song (2016) called the method of divide and combine in regularised generalised linear model for big data. The proposed method is to solve the computational challenges that arising from big data analysis. In general, it is a procedure to subdivide the data recursively into several relatively independent batches, which will be processed alternately in parallel. Then, the results obtained for each batch of data are combined together in a way that algebra and matrix factorization permit [19].

Moreover, a divided regression analysis for big data was carried out by Jun et al., (2015). They performed regression analysis on each sub set iteratively and then got the average result to be compared with the actual result of the whole data set. They also evaluated the proposed model by performing simulation study and checking on the confidence interval of the regression results between divided and full data set. As a result, the regression parameters of the estimated model give slightly difference with the actual model of the entire data set [12].

MATERIALS AND METHODS

Data of the players in the National Basketball Association (NBA) are used to perform the divided regression analysis. Firstly, the entire big data set is divided into n sub data sets with smaller size. The estimated parameter vector $\widehat{\mathbf{B}}$ is defined as $\widehat{\mathbf{B}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k)$ and the estimated regression model is:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1X_1 + \hat{\beta}_2X_2 + \cdots + \hat{\beta}_kX_k$$

In order to build the divided regression model, the $\widehat{\mathbf{B}}_{\boldsymbol{i}}$ where $i = 1,2,\cdots,n$ for each sub data set are used to determine the estimated regression parameter of population, $\widehat{\mathbf{B}}_{\boldsymbol{c}}$ as follow:

$$\widehat{\mathbf{B}}_{\boldsymbol{c}} = f_c(\widehat{\mathrm{B}}_1, \widehat{\mathrm{B}}_2, \dots, \widehat{\mathrm{B}}_n)$$

where f_c is a combine function to combine the results from 1st sub data set to n^{th} sub data set. The mean value is shown as follow:

$$\widehat{\mathbf{B}}_{\boldsymbol{c}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathrm{B}}_i$$

Next, the following step is to evaluate the feasibility of the divided regression model by comparing the mean value of regression parameters of *n* sub data sets with the entire data set.

Besides, variance inflation factor (VIF) is also used to determine multicollinearity between variables. The VIF is computed based on a tolerance, t which is denoted as

$$t = 1 - \mathrm{R}_i^2$$

where R_i^2 is the R-squared value obtained by regression the i^{th} predictor on the remaining predictors. Then the VIF is computed as follow:

$$VIF = \frac{1}{t}$$

A VIF value of 1 means there is no correlation among the independent variables in the regression model, if VIF falls within 4 to 10, indicating further investigation is needed and the VIF that exceeds 10 will lead to serious multicollinearity problem between variables.

There are several assumptions have been used in the study of regression analysis, including linearity of regression model, constancy of error variance and normality of random error [20]. It is needed to consider the validity of these assumptions in the model and analysis is conducted to examine the adequacy of the model built. The validity of these assumptions are studied from residual plots and quantile-quantile plots.

RESULTS AND DISCUSSION

In this study, data of NBA players from year 1997 until year 2017 are used. It was obtained from the website basketball-reference.com. A model is developed to make prediction on the points scored by each NBA player per season.

There are total of 10609 records with four attributes, three independent variables which are FG (percentage of successful field goal), FT (percentage of successful free throw) and MP (minutes played per season) and the dependent variable PTS (total points obtained of player per season).

Statistical results of sub data set

Firstly, the entire set of data is divided into five sub data sets and multiple linear regression is fitted into every sub data set. The multiple linear regression model is formulated as follow:

OPEN ACCESS Freely available online

$$\widehat{PTS} = \hat{\beta}_0 + \hat{\beta}_1 FG + \hat{\beta}_2 FT + \hat{\beta}_3 MP$$

Table 1 shows the values of regression coefficients of each sub data set.

Table 1 NBA player data set result.

Data Set	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Sub 1	-272.1263	120.9114	211.2107	0.45930
Sub 2	-316.9940	159.9039	237.6870	0.4755
Sub 3	-269.1310	97.9386	216.9003	0.4653
Sub 4	-288.5003	155.0568	214.5023	0.4567
Sub 5	-263.4935	107.8175	210.0752	0.4516

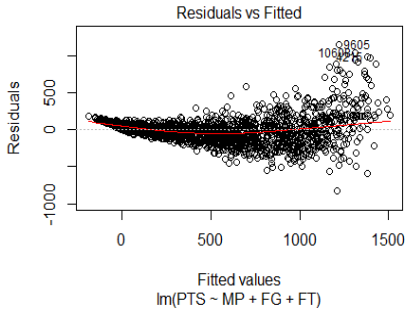

Fig. 1 Residual Plot of Sub Data Set 1.

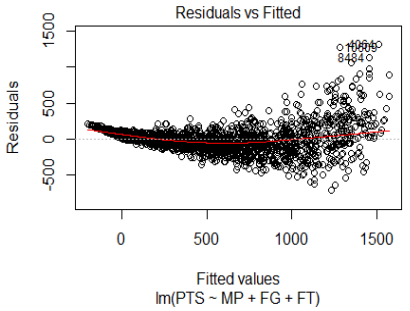

Fig. 2 Residual Plot of Sub Data Set 2.

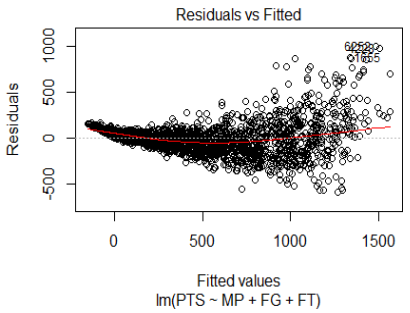

Fig. 3 Residual Plot of Sub Data Set 3.

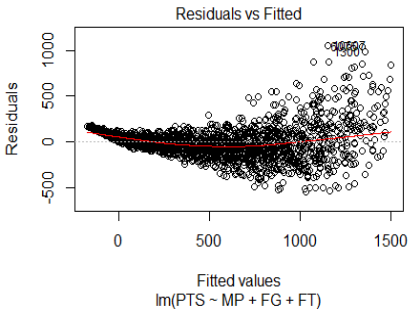

Fig 4 Residual Plot of Sub Data Set 4.

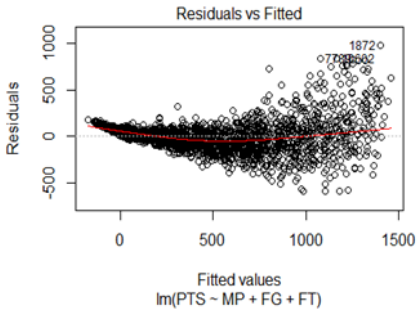

Fig 5 Residual Plot of Sub Data Set 5.

Examine for multicollinearity

Next, multicollinearity between independent variables are checked by using variance inflation factor (VIF). Table 2 shows that all the VIF values are small and closer to 1 and they are not falling into the criteria that having VIF more than 10, indicating the absence of multicollinearity between the independent variables. There are no highly correlated independent variables in the regression model and the multiple linear regression models of the five data sets are said to be safe from multicollinearity.

Table 2 VIF values between independent variables.

Sub data set	Variance Inflation Factor (VIF)		
	FG	FT	MP
1	1.068	1.072	1.109
2	1.112	1.087	1.149
3	1.136	1.105	1.158
4	1.072	1.094	1.133
5	1.117	1.094	1.169

Model adequacy checking

In the study of regression analysis, assumption has been made that the relationship between dependent variable and independent variables is linear. That is, the response variable, PTS should have approximately linear relationship with FG, FT and MP. The linearity of the models are checked by observing the residual plots in Fig. 1-5.

For every sub data set, there is a departure from the linear regression model, violating with the assumption that has been made at the beginning of the analysis. The residual plots indicate that a non linear model is more appropriate for all the sub data sets.

In addition, funnel shape which opens to the right can be observed from the residual plots of the sub data sets. Variability of the residuals increases from the left to the right indicating that the variances of error are not constant. This problem may be caused by the response variable to follow a probability distribution in which the variance is functionally related to the mean. Fig. 1-5 show the residual plots for all sub data sets. From the residual plots, it is clearly shows that the

relationship between the response variables and regressors are not linear for all the sub data sets. Fig. 6-10, it can be observed that the pattern of q-q plots for the five sub data sets look similar, the right end of each plot displays further apart from the straight line. This indicates that there is a problem with the assumption of normality of errors.

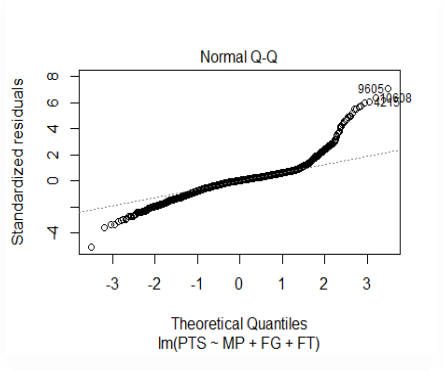

Fig. 6 q-q Plot of Sub Data Set 1.

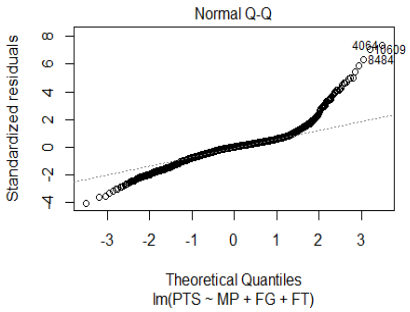

Fig. 7 q-q Plot of Sub Data Set 2.

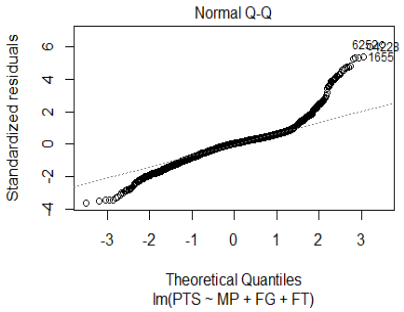

Fig. 8 q-q Plot of Sub Data Set 3.

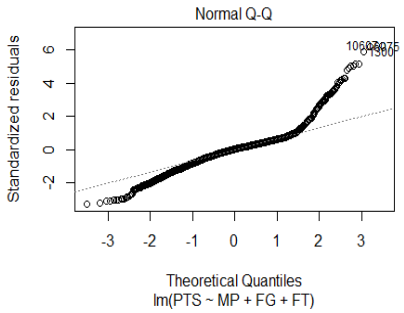

Fig 9. q-q Plot of Sub Data Set 4.

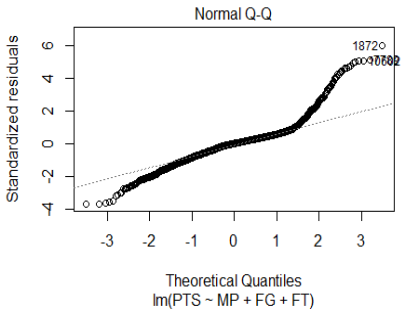

Fig 10. q-q Plot of Sub Data Set 5.

Box-cox method

The results shown that multiple linear regression model is not appropriate for these sub data sets. The transformations on the response variable are attempted, PTS by applying Box-Cox method in order to correct the skewness of the distribution of error terms, non-constancy of error variances and non-linearity of regression model. Table 3 shows the estimated values of $\bar{\lambda}$ for each sub data set, which are obtained from the Box-Cox procedure.

Table 3 Estimated Values of $\bar{\lambda}$ of each sub data set

Sub Data Set	$\bar{\lambda}$
1	0.48
2	0.47
3	0.47
4	0.48
5	0.48

Statistical results and model adequacy checking of new model

Statistical result of the new model for each sub data set are shown in Table 4.

Table 4 NBA Player Data Set Result of new model.

Data Set	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Sub 1	0.64801	4.80039	4.36793	0.00936
Sub 2	-0.36570	6.53548	4.76792	0.00855
Sub 3	-1.29324	8.27804	5.35103	0.00925
Sub 4	-0.59163	6.31721	5.43917	0.00918
Sub 5	-0.42108	6.21015	5.11872	0.00838

Therefore, the new variables in multiple linear regression model of each sub data set becomes:

$$\left(\widehat{PTS}\right)^{\bar{\lambda}} = \hat{\beta}_0 + \hat{\beta}_1 FG + \hat{\beta}_2 FT + \hat{\beta}_3 MP$$

OPEN ACCESS Freely available online

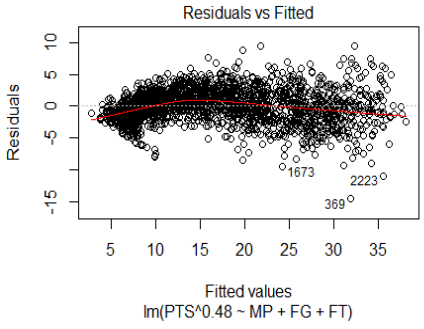

Fig 11. Residual plot of sub data Set 1 after transformation

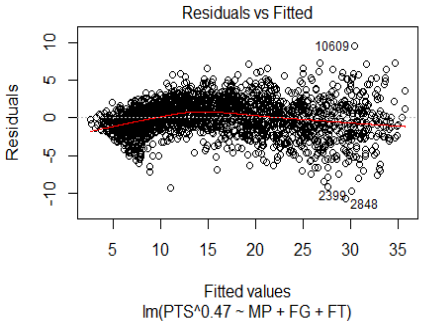

Fig 12. Residual plot of sub data set 2 after transformation

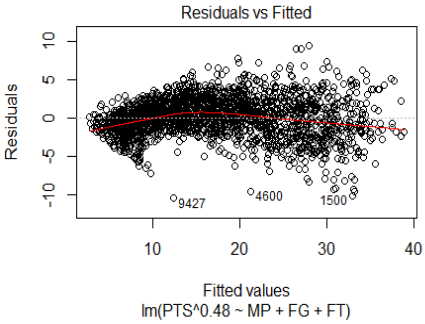

Fig 13. Residual plot of sub data set 3 after transformation

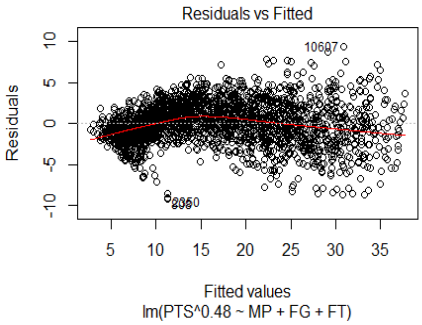

Fig 14. Residual plot of sub data Set 4 after Transformation

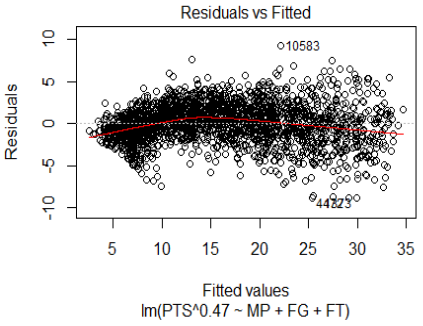

Fig 15. Residual Plot of Sub Data Set 5 after Transformation

From the q-q plots in the Figure 16-20, it has been noticed that the residuals fall within a horizontal band centred around 0. Thus, linearity of multiple regression model is satisfied for every sub data set. Moreover, the funnel shape of residual plots disappears and becomes structureless, indicating that the assumption of constancy of error variances is satisfied in the new variables in the model.

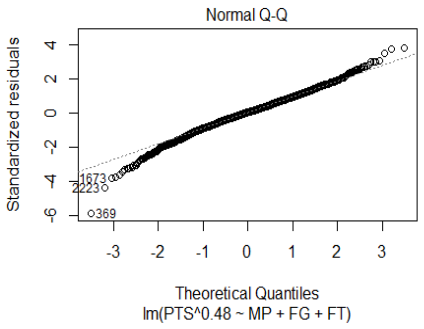

Fig 16. q-q Plot of Sub Data Set 1 after transformation

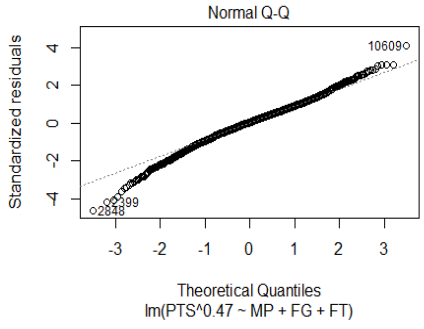

Fig 17. q-q Plot of Sub Data Set 2 after transformation

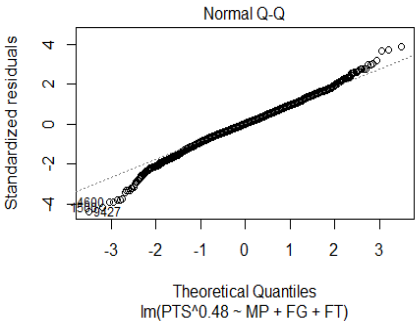

Fig 18. q-q Plot of Sub Data Set 3 after transformation

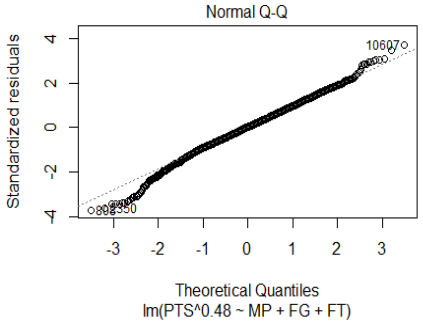

Fig 19. q-q Plot of Sub Data Set 4 after Transformation.

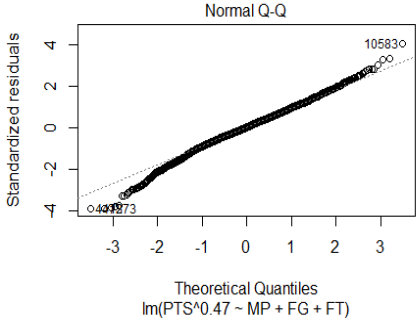

Fig 20. q-q Plot of Sub Data Set 5 after Transformation.

Divided regression model

In order to build the divided regression model for predicting the points scored by each basketball player per season, the mean values of every regression coefficient of all the sub data sets are computed. Then, the values of regression parameters obtained from the mean value function with the actual value of regression parameters of the entire data set are compared. Table 5 shows the values of regression coefficients of each sub data set and entire data set.

Table 5 NBA player data set result for divided regression model.

Data Set	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Sub 1	0.64801	4.80039	4.36793	0.00936
Sub 2	-0.36570	6.53548	4.76792	0.00855
Sub 3	-1.29324	8.27804	5.35103	0.00925
Sub 4	-0.59163	6.31721	5.43917	0.00918
Sub 5	-0.42108	6.21015	5.11872	0.00838
Mean	-0.40473	6.42825	5.00895	0.00894
Entire	-0.47479	6.53451	5.13303	0.00923
Difference	0.07006	0.10626	0.12408	0.00029

From Table 5, it has been noticed that there are differences between the regression coefficients of sub data sets and entire data set, but the mean values of regression coefficients of sub data sets are similar to the values of regression parameters of the entire data set. As the value between mean and entire are very similar, the validity of the divided regression model is confirmed and divided regression model for predicting the performance of NBA player is formulated as follow:

$$PTS = (-0.40473 + 6.42825FG + 5.00895FT + 0.00894MP)^{\frac{1}{0.48}}$$

CONCLUSION

Multiple regression model is used to make prediction on the performance of NBA players based on their field goal rate, free throw rate and minutes played. Instead of applying the normal way of regression analysis, divided regression analysis are proposed in order to reduce computing burden on huge volume of data. This approach divides the entire data set into several sub data sets and multiple linear regression analysis is then performed on every sub data set. Lastly, the results obtained from sub data sets are then combined together to get the divided regression model for the entire data set.

In the procedure of performing analysis on sub data sets, it has been noticed that the assumptions of multiple linear regression model are not satisfied, meaning that multiple linear regression model is not suitable for the data set. Therefore, transformation on the data set by using Box-Cox method are applied in order to make the multiple linear regression model to be appropriate for the transformed data set. Eventually, multiple regression model for each sub data set are obtained and the divided regression model for entire data set is built by taking mean value on the regression coefficients of sub data sets. It has been noticed that there is only slight difference in the values of regression parameters between divided regression model and the model built by performing regression analysis on the entire data set, assuming that it is able to find the actual model of the entire data set. Thus, the performance of divided regression model is verified.

REFERENCES

[1] Draper, N. R., Smith, H. (2014). *Applied regression analysis* (Vol. 326). John Wiley & Sons.

[2] Silhavy, R., Silhavy, P., Prokopova, Z. (2017). Analysis and selection of a regression model for the Use Case Points method using a stepwise approach. *Journal of Systems and Software*, *125*, 1-14.

[3] Stoklosa, J., Huang, Y. H., Furlan, E., Hwang, W. H. (2016). On quadratic logistic regression models when predictor variables are subject to measurement error. *Computational Statistics & Data Analysis*, *95*, 109-121.

[4] Vastrad, C. (2013). Performance analysis of regularized linear regression models for oxazolines and oxazoles derivitive descriptor dataset. *arXiv preprint arXiv:1312.2789*.

[5] Schneider, A., Hommel, G., Blettner, M. (2010). Linear regression analysis: part 14 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, *107*(44), 776-782.

[6] Awang, S. R., Alimin, N. S. N. (2016). The significant factors for the people with epilepsy high employability based on multiple intelligence scores. *Malaysian Journal of Fundamental and Applied Sciences*, *12*(1), 1-5.

[7] Mason, C. H., Perreault Jr, W. D. (1991). Collinearity, power, and interpretation of multiple regression analysis. *Journal of marketing research*, 268-280.

[8] Dubey, R., Gunasekaran, A., Childe, S. J., Wamba, S. F., Papadopoulos, T. (2016). The impact of big data on world-class sustainable manufacturing. *The International Journal of Advanced Manufacturing Technology*, *84*(1-4), 631-645.

[9] Fan, J., Han, F., Liu, H. (2014). Challenges of big data analysis. *National science review*, *1*(2), 293-314.

[10] Gandomi, A., Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144.

[11] Wang, C., Chen, M. H., Schifano, E., Wu, J., Yan, J. (2016). Statistical methods and computing for big data.*Statistics and its interface*, *9*(4), 399.

[12] Jun, S., Ryu, S. J. L. B. (2015). A divided regression analysis for big data. *International Journal of Software Engineering and Its Applications*, *9*(5), 21-32.

[13] Yang, Y. S. (2015). *Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics* (Doctoral dissertation, PhD thesis, UC Berkeley).

[14] Aiken, L. S., West, S. G., Pitts, S. C. 2003. Multiple linear regressions. In, Handbook of Psychology. 19, 481-507.

[15] Fitrianto, A., Hanafi, I., Chui, T. L. (2016). Modeling Asia's Child Mortality Rate: A Thinking of Human Development in Asia. *Procedia Economics and Finance*, *35*, 249-255.

[16] Hall, R., Fienberg, S. E., Nardi, Y. (2011). Secure multiple linear regression based on homomorphic encryption. *Journal of Official Statistics*, *27*(4), 669.

[17] Chen, G. J. (2012). A simple way to deal with multicollinearity. *Journal of Applied Statistics*, *39*(9), 1893-1909.

[18] Chen, X., Xie, M. G. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, 1655-1684.

[19] Tang, L., Zhou, L., Song, P. X. K. (2016). Method of Divide-and-Combine in Regularised Generalised Linear Models for Big Data. *arXiv preprint arXiv:1611.06208*.

[20] Martin, J., de Adana, D. D. R., Asuero, A. G. (2017). Fitting Models to Data: Residual Analysis, a Primer. In *Uncertainty Quantification and Model Calibration*. InTech.

OPEN ACCESS Freely available online