Australian Journal of Basic and Applied Sciences

2020 October; 14(10): pages 1-7
DOI: 10.22587/ajbas.2020.14.10.1

Original paper **AENSI Publications**

The Challenges of Using Students Academic Performances Data and Their Solutions

[1]Keya Rani Das*, [2]Bishnu Kumar Adhikary, [3]Provash Kumar Karmokar

[1]Bangabandhu Sheikh Mujibur Rahman Agricultural University, Department of Statistics, Faculty of Agricultural Economics and Rural Development, Gazipur-1706, Bangladesh.
[2] University of Rajshahi, Institute of Education and Research (IER), Rajshahi- 6205, Bangladesh.
[3] University of Rajshahi, Department of Statistics, Faculty of Science, Rajshahi- 6205, Bangladesh.

Correspondence Author: Keya Rani Das, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Department of Statistics, Faculty of Agricultural Economics and Rural Development, Gazipur-1706, Bangladesh
E-mail:- *keyadas57@bsmrau.edu.bd*

Received date: 22 August 2020, Accepted date: 29 October 2020, Online date: 15 November 2020

Abstract

Background: In educational statistics, quantitative data is getting popular in the world. Although the use of such a data set is increasing day by day in the educational statistics area in Bangladesh, the use of appropriate statistical methodology is strongly recommended.
Objectives: Among the various statistical techniques, somewhat few of them are using correlation and regression analysis from the educational perspective without following the proper methodologies which have been producing misleading output. The current research focuses on such a basic part of the analysis for the proposed educational data set. Specifically, regression analysis uses to check the percentage of variation in the dependent variable explained by variation in the independent variables based on the coefficient of determination (R^2). But R^2 value does not give always a good measure due to the violation of different assumptions. As such R^2 may mislead to judge the influence of independent variables on the dependent variable as a measure. Therefore, in this study, the authors have been intended to focus on the major challenges and its remedies in educational data research from statistical viewpoints.
Results: Simple correlation and regression method have used for the authors' surveyed data following proper assumptions and extreme observations have identified by examining the plot of Cook's distance against centered leverage value. Data analysis completed by IBM SPSS 21.0. The R^2 value computed as 0.792 indicating that there is about 79.2% of total variation explained by the regression line but *VIF* for estimated coefficient b_3 is 10.2. Together with the non-normal pattern of probability plot outlying observations numbered 11, 12, and 35 have identified by scattered diagram between Cook's distance and centered leverage values. The identified extreme observations have removed from the data set and have reexamined. Finally, all sorts of assumptions together with the normality of errors have been confirmed by the reduced data set.
Conclusion: The practitioners need to be careful about the outlying observation involved with their educational data sets so that statistical methodologies would not be applied inappropriately.

Keywords: Educational statistics, Regression, Normality, Multicollinearity, R-Square

INTRODUCTION

Research methods in education and social sciences are mostly divided into two main types which are quantitative and qualitative methods (Muijs, 2011). Statistical tests are using for getting more valid and reliable results. Uses of different types of statistical analysis are spreading out day by day in the field of education of Bangladesh. Regression analysis is a conventional method to analyze data sets and draw inference about the population parameters involved in the model in different research areas. Sometimes educational researchers apply multiple regression incorrect way (Karpen, 2017). As a standard and easy method, its application in educational research is also seen. Regression analysis is one of the most used methods in education in Bangladesh. It was found that most of the researchers had used a regression method for their study directly. To observe the nature of relationship between a dependent variable and one or more independent variables, generally, we use regression analysis. The idea about correlation and regression are available in Pandey (2020), Kumari and Yadav (2018). Not only for educational research but also it is applied as a statistical method widely to conclude the idea or to draw inference about the population based on the sample in the present times. Sometimes the practitioners do not follow the proper ways to apply different statistical methods and techniques. As a result, they do not care about the violation and misinterpretation of the assumptions which may produce an unreliable result. As a common measure for checking the validity or goodness of fit in the regression model, R^2 is used. Some researches reflect that the fitted model is good enough for having its high value of R^2. But for comparing different samples, low value of R^2 cannot give any guarantee about the relationship that it is weak and in the same way, the high value cannot give guarantee about the strong relationship (Achen, 1977). When the linearity assumption breaks, R^2 should not be employed (Kavalseth, 1985). So in this paper, researchers concentrated on identifying the challenges of using primary data influencing students' academic performances within statistical methodologies. This research makes a noteworthy contribution to the current use of the regression method and its proper uses in the field of educational research in Bangladesh. This study also makes open the current situation of the researcher's uses with its challenges of regression method. So, there is a great need of the study to understand whether researchers use regression method properly with its assumption checking or not. Research is, therefore, essential to see what challenges researchers are facing to use regression method in the field of educational research. Regarding the outcome of this article, government, different development organizations, statisticians, education experts, education practitioners, researchers and other concern stakeholders will get a clear idea, as well as they, can take the initial steps to overcome such challenges.

1.1 Research Questions

The main objective of this study is to identify the key challenges of using educational data of students' academic performances in the higher secondary education level of Bangladesh.

The specific research questions are:

a) What are the responsible challenges in analyzing the educational data influencing students' academic performances with basic statistical analysis?
b) What are the suggestions to overcome the challenges of analyzing educational data from a statistical point of view?

2. MATERIALS AND METHODS

Conduct of descriptive and interpretative analysis is more important (Marshall and Rossman, 1999) and the researchers need to pay attention to the data collection technique according to the nature of the study (Creswel, 2012). In this research data have collected from different colleges in Dhaka city based on a structured questionnaire. Firstly, 5 colleges selected and 10 students are taking from each college for an interview. Colleges and students have selected based on Simple Random Sampling (SRS). Researchers have followed the ethical guidelines during data collection and other research activities strictly in this study. Due to limitation of time and budget allocation, the researchers were confined only in Dhaka city for data collection. Data analysis conducted by IBM SPSS 21.0 version.

2.1 Statistical Methods and Diagnostics

2.1.1 Correlation Analysis

Correlation and regression analysis are common and widely used tools in statistical data analyses. Interpretation of correlation and regression results is relatively hard for non-statistician under the theoretical and practical assumptions. The properties, correlation focuses on the cause-and-effect relationship may misguide the researchers (Ludbrook, 2002). Actually, correlation measures the association or strength of linear relationship among two or more variables (Lind at el., 2010).

Karl Pearson (1890) defined the coefficient of correlation as the strength of the relationship between two sets of ratio or interval-scaled variables. Sometimes it is referred to as the Pearson's *r*. Followed by scatter diagram (Bewick et al., 2003) the strength of linear relationship may be measured by the statistic,

$$r_{xy} = \frac{Cov(xy)}{\sqrt{Var(x)Var(y)}}$$

This coefficient of correlation ranges from –1 to 1. In short,

- *r* equals to +1 indicates a perfect positive linear relationship between variables.
- *r* equals to –1 indicates a perfect negative linear relationship between variables.
- *r* equals to 0 indicates no such linear relationship between variables, sometimes it may caused to the existence of the non-linear relationship.

2.1.2 Regression Analysis

Regression measures the relationship of a set of independent variables with the assigned dependent variable. It especially recognizes the number of changes independent due to unit changes of an independent variable(s). The main purpose of regression analysis is to predict and interpret the amount of changes of the response variable in terms of the predictor(s). In this study the academic performance of students is considered as a dependent variable, *y* and the independent variables or explanatory variables are students Study Time, Play Time and TV Watching Time will regress by the linear regression model

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e_i$$

Where Y_i = Academic performance of students; β_0 = the intercept of the line on the Y axis; β_i = the slope of the line indicating the change in Y_i for each unit change in x_i (linear regression co-efficient), i = 1, 2, 3; X_1 = study time; X_2 = play time; X_3 = TV watching time.

The errors, e_i are independently normally distributed with the same variance σ^2 and the assumptions are as-

i. Normality

Residuals are the difference between observed value of dependent variable and the predicted value of the regression. To draw valid conclusion about the data sets verification of normal predicted probability plot (*P-P*) plot is expected where some little dot points will be surrounded to the diagonal line. Although some fluctuations may consider extreme deviation arises due to violet of such assumption will indicate the non-normality of data. When data do not follow normality assumption of regression, then the necessary and sufficient conditions such that the estimates of the parameters are asymptotically normal is difficult (see Huber (1973). In such a case misleading results may produce and the detailed of normality assumption with its test can be found in several studies like, Judge et al. (1985); Das and Imon (2016), etc.

ii. Linearity

Linearity means the straight line relationship in the regression. The relationship between dependent and independent variables need to be linear. To test the linearity, scatter plot is one solution. Also, scatter diagram between residuals and Y values shows the pattern whether the linearity assumption is met or not. When scatter diagram shows a linear pattern, it depicts the assumption is followed. If residuals follow normal distribution as well as homoscedasticity then also the linearity assumption followed. It is also necessary to examine for outliers as outlier effects is a matter for linear regression.

iii. Homoscedasticity

It means the residuals are distributed with equal variances. If data are not homoscedastic, then it turns as heteroscedastic. This assumption can be checked by the scatter diagram of the residuals against the predicted values. If it shows that the dots are distributed uniformly from the zero pointed line then data are homoscedastic.

iv. Multicollinearity

The problem arises due to the high correlation among the independent variables are known as multicollinearity. It is applicable for multiple linear regression where more than one independent variables involves. Variance Inflation Factor (*VIF*) is used to check this assumption. Although the value of *VIF* is less than or equals to 5 is considered as no multicollinearity problem in the dataset, sometimes its level are extended to below 10. A helpful guide about this assumption is presented in Daoud (2017), Kim (2019).

2.1.3 Coefficient of determination (R^2)

Coefficient of determination, R^2 explains the percentage of variation of the dependent variable by the variation in the independent variables (Kennedy, 2008) is defined as

$$R^2 = \frac{Explained\ Variance}{Total\ Variance} \times 100$$

3. RESULTS AND DISCUSSION

In this research correlation and multiple linear regression analysis were employed to know the academic performance of achieving the students GPA (in the scale of 5) as a response to the variables, Study Time (x_1), Play time (x_2), TV watching time (x_3). A brief discussion of the findings followed by authors' surveyed data are given in this section accordingly.

The scatter diagram in Figure 1 shows the irregular behaviour of the data.

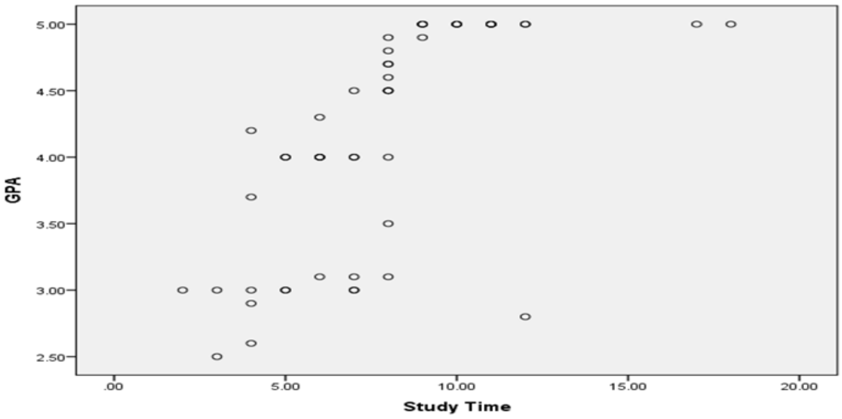

Figure 1: Scatter Diagram between GPA and Study Time

Table. 1: Coefficient of Correlation (r) Values Among GPA, Study Time, Play Time and Spending Time by Watching Television

	GPA	Study Time	Play time	TV Watching Time
GPA	1.000			
Study Time	0.874* (0.01)	1.000		
Play Time	-0.533** (0.000)	-0.080 (0.218)	1.000	
TV Watching Time	-0.949** (0.001)	-0.838** (0.001)	-0.733 (0.464)	1.000

From Table 1, the correlation coefficient between GPA and study time is 0.874, which indicates a positive correlation and p-value is 0.01, which is significant. In the same way, it shows a negative correlation between GPA and playtime and also for GPA and spending time watching television. Also, it discloses a negative relationship between study time and playtime, study time and spending time watching television, playtime and spending time watching television. From this table also we can see independent variables are correlated with each other, which is a sign for multicollinearity.

Table 2: Regression Output and Other Statistics

Subject	**Result**
Constant Value (*a*)	4.7** (0.0001)
b_1 *VIF*	0.084** (0.002) 4.25
b_2 *VIF*	-0.017 (0.200) 1.30
b_3 *VIF*	-0.040** (0.001) 10.20
R^2	0.792
Adjusted R^2	0.688
DW	0.560

Table 2 shows the fitted regression line is academic performance GPA = 4.7 + 0.084*Study time – 0.017*Playtime – 0.040*TV watching time. The R^2 value is also looking high and it may be said that the fitted model is good enough. Some research revealed high R^2 value indicates good model fitting and it finishes in this point. But regression model needs to be checked assumptions whether R^2 value is high or low other than it will be faulty. This is our main issue in this article. Now we want to check some statistic through R^2 value is 0.792. Variance inflation factor (*VIF*) tells us about multicollinearity, which we mentioned above. Multicollinearity can be assessed by *VIF* and the detection of multicollinearity with its exclusion of multicollinear explanatory variables make a good multiple linear regression model (Kim, 2019). From table 2, *VIF* shows the independent variables are correlated, that means multicollinearity arises here. For the regression coefficient (b_3) *VIF* is very high and it is 10.20. This *VIF* for the predictor of spending time by watching television shows us the variance of the estimated coefficient of b_3 is inflated by a factor of 10.20 due to television watching time is highly correlated with at least one of the other independent variables in the model. Now, what should we do? Here the solution is to delete predictor or predictors which are affected by multicollinearity from the model. Now if we have a look on pairwise correlation values from table 1 then we can see that the independent variable of times (in hours) spending by watching television (x_3) and study time (x_1) are highly correlated ($r = –0.838$). We can choose any independent variables and remove one of them x_1 or x_3 from the model. It depends on the researcher and the purpose of research. In this case, study time is more important than television watching time for students result. So researcher can omit x_3 variable and needs to fit a regression model again with other independent variables. Then it will be more reliable model in statistical perspective.

According to Durbin Watson *DW* statistics gives us the idea of autocorrelation in the residuals from a statistical regression analysis. Its range is 0 to 4 and the value of 2 indicates that there is no autocorrelation. In this case, *DW* is 0.560 which focuses autocorrelation present here.

Together with these characteristics, one more test, Dickey Fuller (*DF*) have been used to check the stationarity of the data set. By such stationarity test the null hypothesis, the data is non-stationary against the stationary of the data have used following the *DF* test statistic. The *DF* test statistics for y, x_1, x_2 and x_3 are –2.8041, –5.5849, –2.9913 and –4.2959 respectively. The critical value of the tests are –2.57, indicating that the hypothesis, *the data is non stationary* is accepted for all the variables indicating that all the data sets are non-stationary.

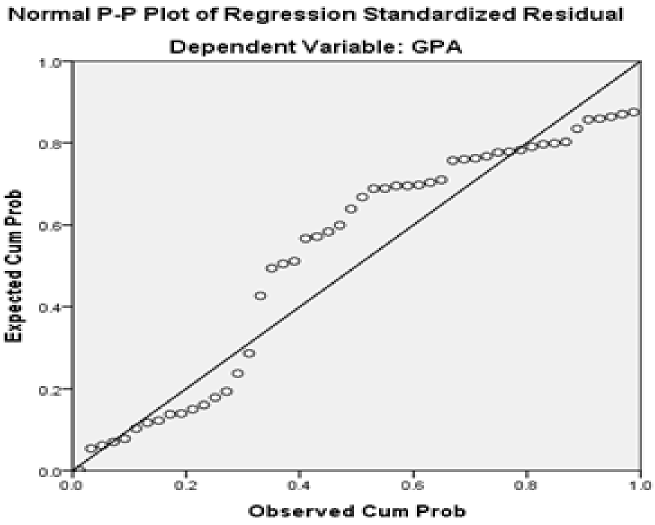

Figure 2: Normal Probability Plot

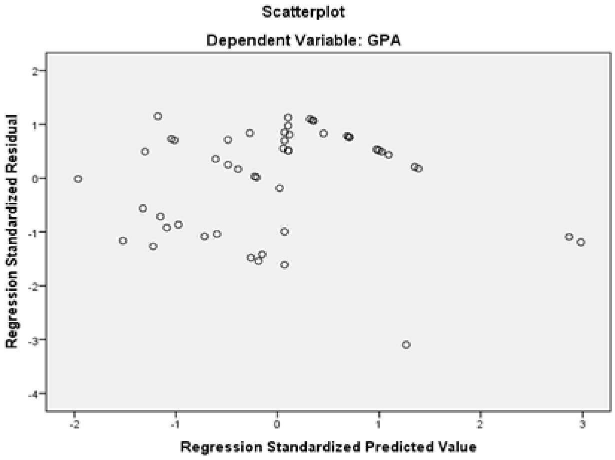

Figure 3: Scatter Diagram between Residuals and Predicted Values

Figure 2 shows the violation of normality for the data. Also, the scatter diagram of residuals and predicted values (in Figure 3) shows some heteroscedastic pattern for residuals. As such, there may be some extreme values in the data set. Observations are statistical outliers that deviate abnormally from the overall shape of data (Jones, 2019). The plot of Cook's Distance against Centered Leverage Value given in Figure 4, indicating having outliers in the data set. The data in the position 11, 12 and 35 are identified as the extreme/outlying observations.

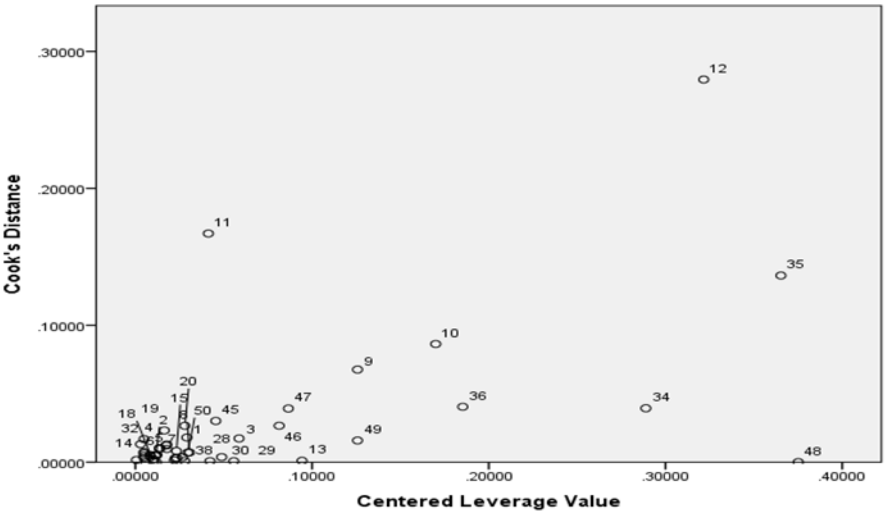

Figure 4: Plot of Cook's Distance against Centered Leverage Value

Now, after deleting extreme values (serial no. is 11, 12 and 35) again, we fit a multiple regression line. The following table gives coefficient values and R^2.

Table 3: Regression Output after Deleting Outlier

Statistic	Result
Constant (*a*)	2.898** (0.0002)
b_1 *VIF*	0.165** (0.001) 1.003
b_2 *VIF*	-0.020** (0.0004) 1.02
R^2	0.780
DW	1.720

Finally the fitted regression line is GPA (y) = 2.898 + 0.165 (Study time) – 0.020 (Play time). Here R^2 value is also high but slightly differ from observation with extreme values. Again Durbin Watson (*DW*) statistic is near to 2, which is an indication of a good result. Also if we compare the normal probability plot of outliers in figure 1 with removing outliers in figure 5, which shows that data without outliers follows normal distribution which is a very common and important condition. If we check assumptions carefully and take the necessary steps by controlling statistical diagnosis, then it will be more accurate.

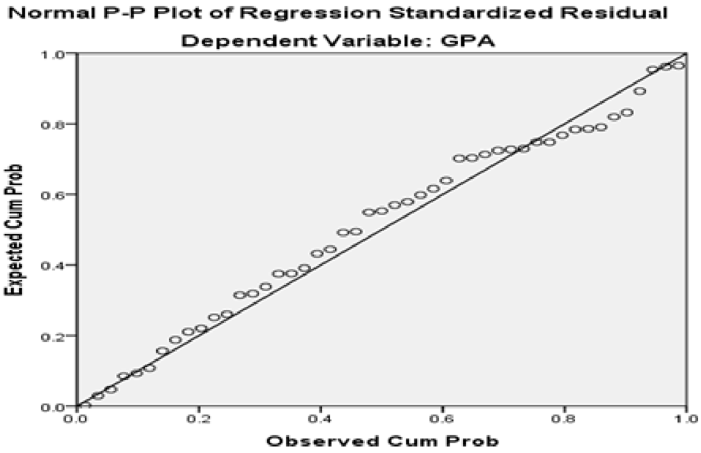

Figure 5: Normal Probability Plot after Deleting Outliers

4. CONCLUSION

This study aims to identify the challenges of using primary data of students' academic performances based on the proper application of statistical techniques. From the statistical viewpoint, some simple mistakes and misuses in regression analysis are noticed following primary data in educational research. As the non-stationary problem infects the data set, it is mandatory to check the necessary diagnostic tools before confirming the fitted model. As faulty and unreliable results may happen without ensuring the recommended assumptions properly, the researchers focus on such problems following its step by step solutions.

Since the data possess a severe outlying problem, the main challenge of the current research is to draw a substantial conclusion. Hence the extreme values (serial no. is 11, 12, 35) in the data set have discovered with the help of *Cook's Distance* and *Centered Leverage Value and VIF*. Finally, the R^2 diagnostics, *DW* statistics have been re6calculated with the help of modified data set where identified outliers were deleted. All the problems involved with the data set are now dissolved, which have confirmed by the existence of non-autocorrelation and the normality of errors. Therefore, the findings are now reliable and a message of knowing the potentiality and limitations of methodological choice is remarkable for the practitioners.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Achen, C., H.,1977. Measuring Representation: Perils of the Correlation Coefficient. *American Journal of Political Science*. Vol. 21:805-815. DOI: 10.2307/2110737

Bewick, V., Cheek, L., and Ball, J, 2003. Statistics review 7: Correlation and regression. *Critical Care.*Vol. 7(6):451–459. DOI: org/10.1186/cc2401

Creswell, J. W, 2012.. *Educational research: Planning, conducting, and evaluating quantitative and qualitative research (4th ed.).* Boston, MA: Pearson.

Daoud, J. I., 2017. Multicollinearity and Regression Analysis. *Journal of Physics : Conference Series.* Vol. 949(1): 1-6. DOI: 10.1088/1742-6596/949/1/012009

Das, K. R., and Imon, A. H. M. R., 2016. A Brief Review of Tests for Normality. *American Journal of Theoretical and Applied Statistics.* Vol. 5(1): 5-12. DOI:10.11648/j.ajtas.20160501.12.

Huber, P.J., 1973. Robust regression: Asymptotics, conjectures, and Monte Carlo. *The Annals of Statistics.* Vol. 1(5):799-821. DOI: 10.1214/aos/1176342503.

Jones, P. R., 2019. A note on detecting statistical outliers in psychophysical data. *Atten Percept Psychophys.* Vol. 81: 1189–1196. https://doi.org/10.3758/s13414-019-01726-3

Judge, G.G., Griffith, W.E., Hill, R.C., Lutkepohl, H., and Lee, T., 1985. “Theory and Practice of Econometrics,” *Wiley,* New York, 2nd.Ed.

Karpen, S. C., 2017. Misuses of Regression and ANCOVA in Educational Research. American Journal of Pharmaceutical Education. Vol. 81(8): 6501. DOI: 10.5688/ajpe6501

Kennedy, P., 2008. “A Guide to Econometrics,” San Francisco, CA: Wiley-Blackwell. Pp-14. 6thedition.

Kim, J. H., 2019. Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology.* Vol. 72(6): 558-569. DOI: 10.4097/kja.19087

Kumari, K. and Yadav, S., 2018. Linear regression analysis study. *Journal of the Practice of Cardiovascular Sciences.* Vol. 4(1): 33-36.

Kvalseth, T., O., 1985. Cautionary Note About R^2. *The American Statistician.* Vol. 39: 279-285. DOI:https://doi.org/10.1080/00031305.1985.10479448

Lind, D. A., Marchal, W. G. and Wathen, S. A., 2010. “Statistical Techniques in Business & Economics,” McGraw-Hill Irwin. Pp 467-478.

Ludbrook, J., 2002. Statistical Techniques for Comparing Measurers And Methods Of Measurement: A Critical Review. *Clinical and Experimental Pharmacology and Phisiology.* Wiley Online. 29(7):527-536. DOI: https://doi.org/10.1046/j.1440-1681.2002.03686.x

Marshall, C and Rossman, G. B., 1999. *Designing Qualitative Research*, Third edition, Thousand Oaks, CA: SAGE publications.

Muijs, D., 2011. *Doing Qualitative Research in Education with SPSS,* Second edition, London: SAGE publications.

Pandey, S., 2020. Principles of correlation and regression analysis. *Journal of the Practice of Cardiovascular Sciences.* Vol. 6(1): 7-11. DOI: 10.4103/jpcs.jpcs_2_20