Australian Journal of Basic and Applied Sciences

2021March; 15(3): pages 16-23
https://doi.org/10.22587/ajbas.2021.15.3.3

REVIEW PAPER AENSI PUBLICATIONS

# How to Conduct Correlation and Regression Analysis through SPSS

**[1]Keya Rani Das*, [2]Khandoker Saif Uddin**

[1]Bangabandhu Sheikh Mujibur Rahman Agricultural University, Department of Statistics, Faculty of Agricultural Economics and Rural Development, Gazipur-1706, Bangladesh
[2]International University of Business Agriculture and Technology (IUBAT), Department of Statistics, Uttara, Dhaka 1230, 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.
Email: keyadas57@bsmrau.edu.bd

**Received date: 28 December 2020, Accepted date: 28 March 2021**

**Cite as:** Das, Keya Rani., & Uddin, K. S., 2021. How to Conduct Correlation and Regression Analysis through SPSS. Australian Journal of Basic and Applied Sciences, 15(3): 16-23. DOI: 10.22587/ajbas.2021.15.3.3.

**ABSTRACT**

**BACKGROUND: Statistics has played its role in everywhere as each field of science, agricultural research, medical and engineering research, social science research, business and economics field and so on. In each field, data should be analyzed and software makes it easy to handle to analyze the data. Among many software SPSS is simple to use.**
**OBJECTIVES: For applied research and test of hypothesis somewhere we cannot go further without touch statistical task. But sometimes we face problems when we conduct basic research by using software as we don't know how to run those software. That's why this paper mainly focused on how to conduct correlation and regression analysis by statistical software named SPSS.**
**FINDINGS: IBM SPSS (Statistical Package for the Social Sciences) 21 version software was used and screenshots from this are listed here to focus how we can conduct correlation and regression analysis by SPSS. In this article a step by step analyzing process about correlation and regression are presented for collected data. This paper also showed how to check assumptions to fit regression line through SPSS. Also some basic idea about correlation and regression are listed here.**
**CONCLUSION: Fundamentals of correlation and regression are listed and also stepwise solution to operate SPSS is featured here about these basic statistical analyses. This paper will help to make familiar and easy to handle SPSS on correlation and regression analysis**

**Keywords:Statistics, Correlation, Regression, SPSS**

## INTRODUCTION

Correlation and regression are the most common terms in statistics. In many types of research, these statistical methods are applied. The SPSS is a very user-friendly software and easy to conduct statistical analysis. The elaboration of SPSS is the Statistical Package for the Social Sciences (Muijs, 2011). So many statistical analyses can be performed through SPSS (Arkkelin, 2014). This software can present statistical analysis and graphical presentation from a variety of data. Descriptive statistics like frequency distributions, measures of central tendency, measures of dispersion, plots, charts, and inferential statistics like tests and multivariate analysis like factorial analysis, cluster analysis, and categorical data analysis can be performed through SPSS. Firstly a question can arise as that why does this SPSS software is chosen for data analysis.

The main features of SPSS are that it is user-friendly and easy to learn. It covers in-depth statistical tools for analysis, and also graphical presentation can be used easily through this software. Also, it is noted that SPSS provides a data management system with its editing tools. From research, the result shows that only 19.43% and 33.18% of students have a clear idea and answered a

satisfactory explanation about correlation and regression coefficient, respectively (Razak et al., 2017). So the basic application of correlation and regression needs to be understandable to the students. In this point, this article also focuses on correlation and regression analyses using SPSS.

This study contributes to the present state of applying the regression method correctly by using SPSS. This paper aims to focus on analysing data with correlation and regression techniques by using SPSS software. Simultaneously, a step-by-step solution to conduct correlation and regression analysis using SPSS must help different research purposes. In this regard, this paper aimed to present various commands along with the screenshots for SPSS software.

## METHODS

In this study, the data have been collected from the students of the University of Rajshahi in Bangladesh through a newly developed questionnaire. The study area was selected purposively. Thirty students enrolled to take face-to-face interviews. During the data collection, ethical guidelines are followed by the researcher strictly. Because of time limitations and minimizing cost, the researchers were confined only to the University of Rajshahi for data collection.

### Statistical Methods

### Correlation and Regression

Correlation measures the strength of the linear relationship. It does not explain the cause-and-effect relation (Oster and Enders, 2018). Correlation coefficient (r) presents the degree of linear association between two variables (Taylor, 1990). The Scatter diagram shows the linear or non-linear pattern of data (Subrata and Das, 2018).

Correlation coefficient r can be defined as

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

Correlation presents the strength and direction of a linear relationship between two variables, whereas regression presents the nature of the relationship by a mathematical equation (Das et al., 2020a). Das et al. ( 2020) and Das et al. ( 2020a) have shown the necessary correlation and regression analysis steps.

Mathematically, the linear regression model is

$$Y_i = \alpha + \beta X_i + e_i;\ \ i = 1, 2, \ldots\ldots n$$

Where $\alpha$ = the intercept term of the line on the Y-axis.

$\beta$ = Regression coefficient or slope of the line indicating the change in the dependent variable for each unit change in the independent variable.

$y_i$ = Dependent variable.

$x_i$ = Independent variable.

Here $e_i$'s are independently normally distributed with the same variance $\sigma^2$.

We should check some assumptions properly to apply the linear regression model, which is as follows:

a. Normality
b. Linearity
c. Homoscedasticity
d. Multicollinearity
e. Outliers
f.

Description: These assumptions are presented in (Das et al., 2020; Das et al., 2020a; Montgomery, 1982).

### Correlation and regression analysis in SPSS

### To draw a scatter plot

The commands are

Graphs → Legacy Dialogs → Scatter/Dot → Simple Scatter → Define → Move the variables into X-axis and Y-axis → OK

### To draw a matrix scatter plot

The commands are

Graphs → Legacy Dialogs → Scatter/Dot → Matrix scatter → Define → Move the variables into Matrix Variables → OK

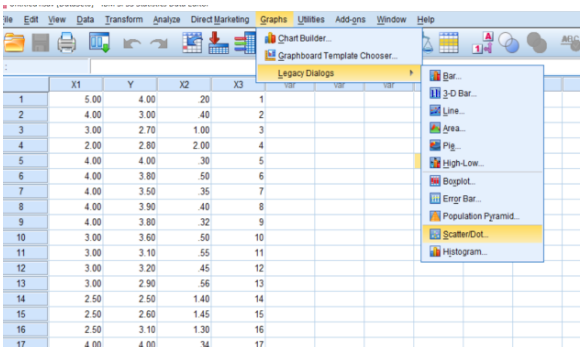


**Figure 1:** The first step to draw scatter diagram

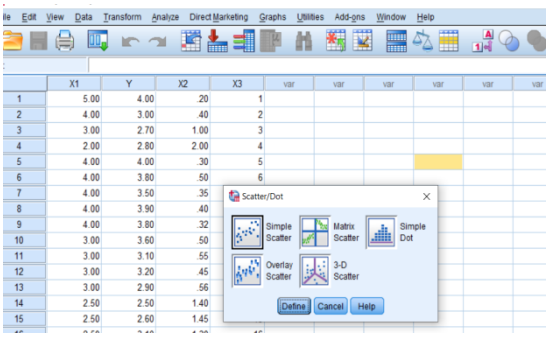


**Figure 2:** Steps of drawing matrix scatter plot

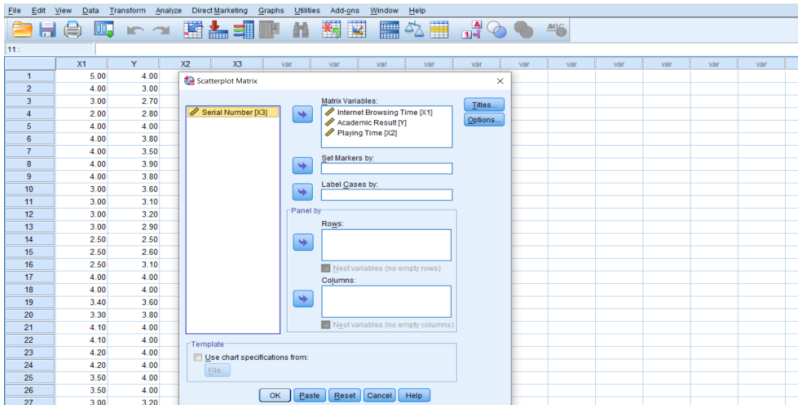


**Figure 3:**Steps of matrix scatter plot drawing

To calculate the coefficient of correlation

Analyze → Correlate → Bivariate → Move variables into Variables box → For correlation Coefficient, make sure that Pearson is being selected→ OK

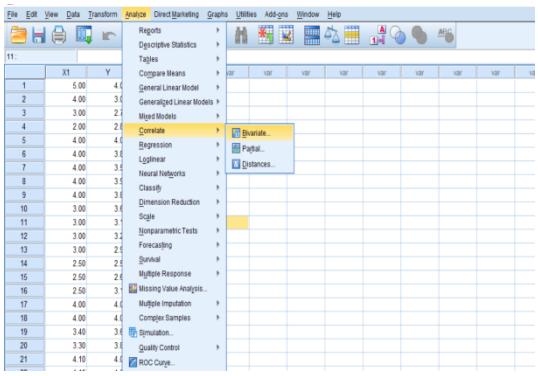


**Figure 4:**First step to calculate r in SPSS

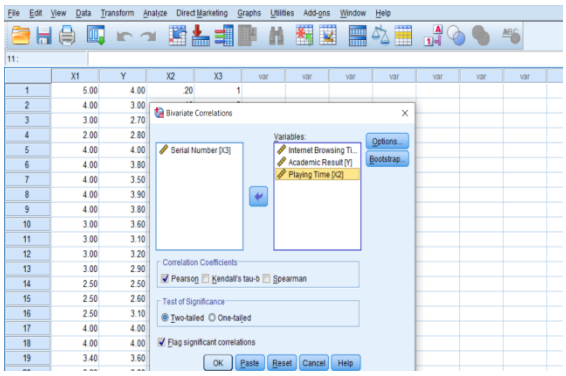


**Figure 5:**Steps to find r value in SPSS

For regression analysis

Analyze → Regression → Linear → Move variables according to dependent and independent → Statistics (From Window) → Chose Estimates, Model fit → Continue → OK.

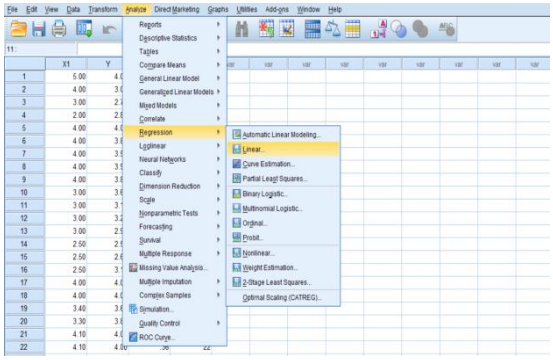

**Figure 6:**First step for regression analysis

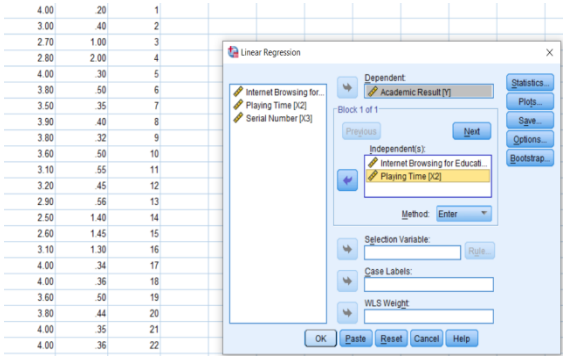

**Figure 7:**Steps for regression analysis in SPSS

**To draw a scatter plot with the regression equation**

Graphs → Legacy Dialogs → Scatter/Dot → Simple Scatter → Define → Move the variables into X-axis and Y-axis → OK
In scatter plot, right-click and choose Edit Content → In Separate Window → Elements → Fit Line at Total**.**

Again, Analyze → Regression → Linear →Move variables according to dependent and independent → Plots (in sub-dialogue box) → Select *ZRESID and move it into Y-axis box →Select *ZPRED and move it into X-axis box → Select Normal Probability Plot (standardized Residual Plots box) → Continue → OK.

**Commands for multiple linear regression with multicollinearity checking**

Analyze → Regression → Linear →Move dependent variable into the Dependent box → Move independent variables into the Independent box→ From Method make sure that Enter is selected → Statistics (From Window) → Chose Estimates, Model fit, Collinearity diagnostics → Continue → OK.

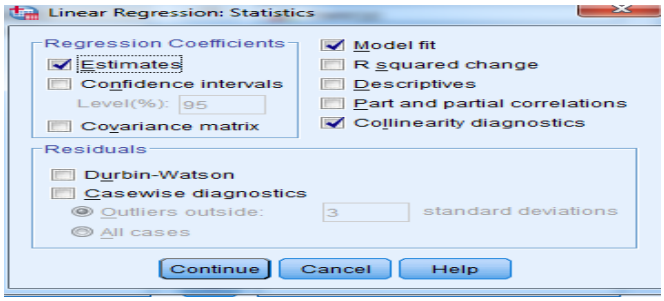


**Figure 8:** Multicollinearity diagnostics in SPSS

**To detect outliers**

Analyze → Regression → Linear →Move dependent variable into the Dependent box → Move independent variables into the Independent box→ From Method make sure that Enter is being selected → Save (From Window) → Cook's Distance (Values will be saved in the data file as variable labeled "COO-1") → OK.

Then need to perform boxplot with this variable "COO-1". The commands are given below to perform boxplot
Graphs → Boxplot → Chose Simple → Select Summaries of Separate Variables → Pass "COO-1" into Boxes Represent → Chose serial number to identify the cases in the "Label Cases By" box → OK.

## RESULTS AND DISCUSSION

After giving a statistical procedure in SPSS, an Output Viewer window is opened where the results remain (Landau and Everitt, 2004). After applying all these commands, the output screen shows the following plot and seems to the linear relationship between variables. Now we can apply correlation analysis for these variables. A data set is attached in the appendix 1 section which is used here to see the results from SPSS.

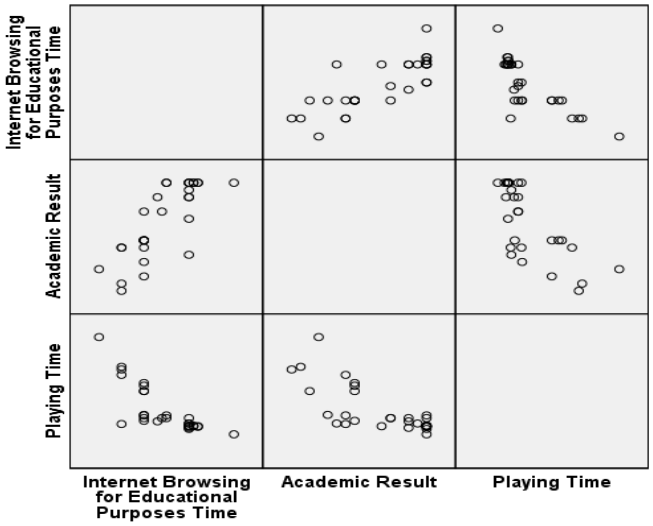


**Figure 9:** Matrix scatter plot

From this matrix scatter plot, there is a linear relation, and the relationship is strongly positive between academic result and internet browsing time for educational purposes. Again, there is a negative correlation between theoretical results and playing time. Also, table 1 presents the *Pearson* correlation coefficient values among variables.

**Table 1:** Correlations

| | | Academic Result | Internet Browsing for Educational Purposes Time | Playing Time |
|---|---|---|---|---|
| Academic Result | Pearson Correlation | 1 | .801** | -.737** |
| | Sig. (2-tailed) | | .000 | .000 |
| | N | 30 | 30 | 30 |
| Internet Browsing for Educational Purposes Time | Pearson Correlation | .801** | 1 | -.790** |
| | Sig. (2-tailed) | .000 | | .000 |
| | N | 30 | 30 | 30 |
| Playing Time | Pearson Correlation | -.737** | -.790** | 1 |
| | Sig. (2-tailed) | .000 | .000 | |
| | N | 30 | 30 | 30 |

**. Correlation is significant at the 0.01 level (2-tailed).

The Pearson correlation coefficient between the academic result (in CGPA) and internet browsing time for educational purposes is .801 with a significance level is 0.000. So there is a strong positive correlation between these two variables. Again, the r-value between the academic result (in CGPA) and playing time is -0.737 and it is also highly significant. There is a strong negative correlation between these two variables.

Regression outputs are as follows

**Table 2:** Model Summary[b]

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .818[a] | .670 | .645 | .29944 |

a. Predictors: (Constant), Playing Time, Internet Browsing for Educational Purposes Time
b. Dependent Variable: Academic Result

Here Model Summary box gives an idea about the proportion of the variation of the dependent variable explained by the independent variable by R Square value. Here it is .670, which means independent variables can explain 67% variation.

**Table 3:**ANOVA[a]

| Model | | Sum of Squares | Df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 4.914 | 2 | 2.457 | 27.401 | .000[b] |
| | Residual | 2.421 | 27 | .090 | | |
| | Total | 7.335 | 29 | | | |

a. Dependent Variable: Academic Result
b. Predictors: (Constant), Playing Time, Internet Browsing for Educational Purposes Time

**Table 4:**Coefficients[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | T | Sig. | Collinearity Statistics | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Tolerance | VIF |
| 1 | (Constant) | 2.252 | .562 | | 4.009 | .000 | | |
| | Internet Browsing for Educational Purposes Time | .418 | .130 | .581 | 3.220 | .003 | .375 | 2.665 |
| | Playing Time | -.318 | .206 | -.278 | -1.539 | .135 | .375 | 2.665 |
| a. Dependent Variable: Academic Result | | | | | | | | |

This table shows the regression coefficients result and the test and its significance level regression coefficient. The first arrow depicts the value of the coefficients and the second arrow displays the t value for the test of the slope of the regression line. The next column (Sig.) shows the significance level for this t-test. For this example data, the fitted regression line is
Academic result = 2.252 + 0.418 (internet browsing time for educational purposes) – 0.318 (playing time).

To check assumptions, by applying commands, the following graphs show in the output screen.

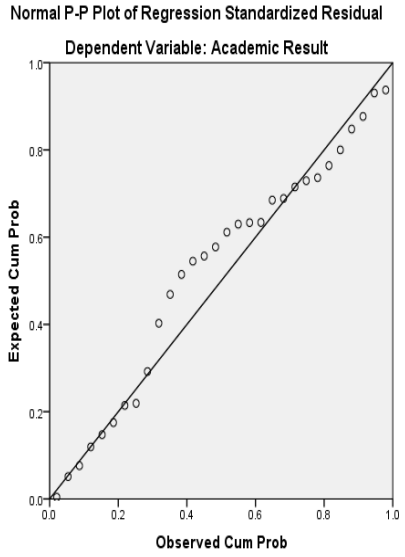


**Figure 10:**Normal P-P plot

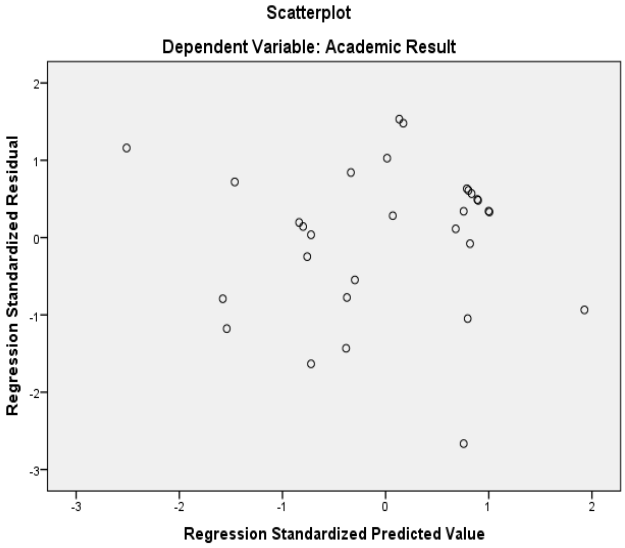


**Figure 11:** Scatter plot between residuals and predicted values

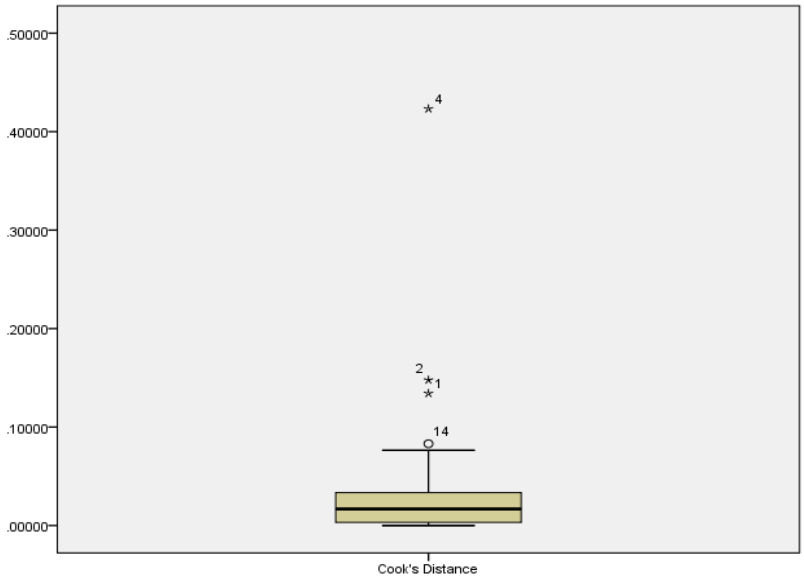


**Figure 12:** Boxplot of Cook's Distance

The box plot showed an outlying problem in the data and remarked the serial number by a star in the graph. Observations that deviate abnormally among overall data shapes treated as statistical outliers (Jones, 2019). Some researchers suggested removing the outlying cases, and some researchers suggested using a nonparametric or alternative regression method when data was affected by outlying problems (Das et al., 2020a). Also, issues and solutions are presented in Das et al., 2020 to estimate the regression model in educational data.

## CONCLUSION

Correlation and regression analysis are prevalent tools in the field of statistics. Statisticians and practitioners are used to applying these tools in different researches. A step-by-step solution for correlation and regression analysis in collected educational data using IBM SPSS is presented in this study. This paper also helps to get a clear idea about how we can check assumptions to fit a regression model using SPSS. The limitations of this study are that data collected from only the University of Rajshahi in Bangladesh. The paper aimed to focus on how to deal with correlation and regression analysis by using SPSS, so data was collected in a limited area due to time and cost limitations.

## REFERENCES

Arkkelin, D., 2014. Using SPSS to Understand Research and Data Analysis. Valparaiso: Valparaiso University.

Das, K. R., B. K. Adhikary, and P. K. Karmokar, 2020. The Challenges of Using Students Academic Performances Data and Their Solutions. Australian Journal of Basic and Applied Sciences, 14(10): 1-7. DOI: 10.22587/ajbas.2020.14.10.1.

Das, K. R., B. K. Adhikary, and K. S. Uddin, 2020a. The Relationship between Students' Results and Spending Time on Internet at Higher Education Level in Bangladesh. ULAB Journal of science and engineering, 11(1): 1-8.

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

Landau, S., and B. S. Everitt, 2004. A Handbook of Statistical Analyses using SPSS. CHAPMAN & HALL/CRC.

Montgomery, D., 1982. Introduction to linear Regression Analysis. New Delhi: Willy.

Muijs, D., 2011. Doing quantitative research in education with SPSS. London: SAGE Publications Ltd. DOI: 10.4135/9781849203241.

Oster, R. A., and F. T. Enders, 2018. The Importance of Statistical Competencies for Medical Research Learners. Journal of Statistics Education, 26(2): 137-142. DOI: 10.1080/10691898.2018.1484674.

Razak, F. A., N. Baharun, N. A. Deraman, and N. R. P. Ismail, 2017. Assessing Students' Abilities in Interpreting the Correlation and Regression Analysis. Journal of Fundamental and Applied Sciences, 9(5S): 644-661. DOI: http://dx.doi.org/10.4314/jfas.v9i5s.45.

Subrata, N., and P. Das, 2018. Applications of different Statistical Tests in Educational Reserarch: An Overview. Journal of Emerging Technologies and Innovative Research, 5(5): 129–137.

Taylor, R., 1990. Interpretation of the Correlation Coefficient: A Basic Review. Journal of Diagnostic Medical Sonography, 6(1): 35-39. doi:10.1177/875647939000600106.

## APPENDIX 1

Data sheet

| Academic result (in the scale of 4.00) | Internet Browsing time for Educational Purposes (in hours) | Playing time (in hours) |
|---|---|---|
| 4.00 | 5.00 | .20 |
| 3.00 | 4.00 | .40 |
| 2.70 | 3.00 | 1.00 |
| 2.80 | 2.00 | 2.00 |
| 4.00 | 4.00 | .30 |
| 3.80 | 4.00 | .50 |
| 3.50 | 4.00 | .35 |
| 3.90 | 4.00 | .40 |
| 3.80 | 4.00 | .32 |
| 3.60 | 3.00 | .50 |
| 3.10 | 3.00 | .55 |
| 3.20 | 3.00 | .45 |
| 2.90 | 3.00 | .56 |
| 2.50 | 2.50 | 1.40 |
| 2.60 | 2.50 | 1.45 |
| 3.10 | 2.50 | 1.30 |
| 4.00 | 4.00 | .34 |
| 4.00 | 4.00 | .36 |
| 3.60 | 3.40 | .50 |
| 3.80 | 3.30 | .44 |
| 4.00 | 4.10 | .35 |
| 4.00 | 4.10 | .36 |
| 4.00 | 4.20 | .34 |
| 4.00 | 4.20 | .35 |
| 4.00 | 3.50 | .50 |
| 4.00 | 3.50 | .55 |
| 3.20 | 3.00 | 1.00 |
| 3.20 | 3.00 | 1.10 |
| 3.20 | 3.00 | 1.15 |
| 3.10 | 2.50 | .39 |