

Comparative Analysis of Five Statistical Packages' Features and Output

Oluwakemi Sade Ayodele & Alabi Taiye John

Department of Computer Science

Kogi State Polytechnic

Lokoja, Kogi State, Nigeria

E-mail: Kemtemmy2009@gmail.com

Phone: +234806 980 4373

ABSTRACT

In this new era of 'big data', the use of statistical software has become inevitable and choosing the right data analysis software is becoming an important aspect of research in virtually any field of human endeavour. Statistical packages are collections of software designed to aid in statistical analysis and data exploration. The vast majority of quantitative and statistical analysis relies upon statistical packages for its execution. An understanding of statistical packages is very essential. Statistical analysis can be implemented using programming languages like C++, Java, and FORTRAN e.t.c but statistical packages are time, effort and cost saving also providing a common interface for data manipulation, visualization and statistical analysis. However, statistical packages vary greatly with respect to accuracy and reliability and reported results may be dependent on the specific package and version. The aim of this paper therefore is to compare the features and output of regression analysis of five statistical packages: Microsoft Excel (2007), R package, SPSS, GraphPad and Stata. In this paper, regression analysis was used to determine the intercept and coefficients of a relation and the results obtained were tabulated. Therefore, for solution that requires high accuracy with speed, Microsoft excel is highly recommended.

Keywords: Data manipulation, Visualization, Data Exploration, Statistical Analysis

iSTEAMS Proceedings Reference Format

Oluwakemi Sade Ayodele & Alabi Taiye John (2019): Comparative Analysis of Five Statistical Packages' Features and Output.

Proceedings of the 19th iSTEAMS Multidisciplinary Conference, The Federal Polytechnic, Offa, Kwara State, Nigeria. 7th – 9th August, 2019.

Pp 115-124. www.isteam.net/offa2019 - DOI Affix - <https://doi.org/10.22624/AIMS/iSTEAMS-2019/V19N1P15>

1. INTRODUCTION

A wide range of software statistical packages can be used to analyse data. These ranges from Access or Excel to dedicated packages such as SPSS, Stata and R for **statistical** analysis of quantitative data, Nvivo for **qualitative** (textual and audio-visual) data analysis (QDA), or ArcGIS for analysing **geospatial** data. In this paper, emphasis is made on the quantitative data Analysis. The five statistical packages used in this paper are **Statistical Packages for Social Sciences** (SPSS), R Package, GraphPad, Microsoft Excel and Statistics/Data Analysis (Stata).

Cavaliere (2015) observed that we live in "data era" where the use of statistical or data analysis software is inevitable in any research field. This means that the choice of the right software tool or platform is a strategic issue for a research department. Nevertheless, in many cases users of statistical software do not pay the right attention to a comprehensive and appropriate evaluation of what the intended use of the result of the data analysis is. Indeed, the choice still depends on few factors like, for instance, researcher's personal inclination, e.g., which software is already known, which shouldn't be the case.

According to Godsey (2019), it's often helpful if a statistical tool can perform some related methods. Often, you'll find that the method you chose doesn't quite work as well as you'd hoped, and what you learned in the process leads you to believe that a different method might work better. If your software tool doesn't have any alternatives, then you're either stuck with the first choice or you'll have to switch to another tool. This paper therefore is an eye opener to choosing the right statistical packages for regression analysis.

2. METHODOLOGY AND DATA PRESENTATION

In this paper, secondary data are used. The data for analysis is the results of an experiment on the impact of data size, execution time and power on energy consumption of sorting algorithm. (Ayodele & Oluwade, 2019). Below is the table of Quick Sort Algorithm implementation in C programming language used for this work.

Table 3.1: Quick Sort Algorithm Implementation In C

Data Size	Average Execution Time(Sec)	Power (Watt)	Energy (Joule)
100,000	0.0594	2.42	0.143748
200,000	0.125	1.9	0.2375
300,000	0.1372	3.54	0.485688
400,000	0.2652	2.56	0.678912
500,000	0.3902	3.4	1.32668

Statistical Packages

The following five Statistical packages are selected for comparison:

Microsoft Excel: This is part of the Microsoft Office suite of programs. Excel version 1.0 was first released in 1985, with the latest version Excel 2016. (Michael Lewis-Beck, 2004)

SPSS: SPSS stands for Statistical Package for the Social Sciences. It was one of the earliest statistical packages with Version 1 being released in 1968, well before the advent of desktop computers. It is now on Version 23.

R Package: This is a collection of R functions, complied code and sample data. They are stored under a directory called "library" in the R environment. By default, R installs a set of packages during installation. More packages are added later, when they are needed for some specific purpose. When we start the R console, only the default packages are available by default. Other packages which are already installed have to be loaded explicitly to be used by the R program that is going to use them. R is a free version of S-plus developed in 1996. Since then the original team has expanded to include dozens of individuals from all over the globe. (https://www.tutorialspoint.com/r/r_packages.htm)

GraphPad: This is a commercial scientific 2D graphing and statistics software available for both Windows and Macintosh computers (https://en.wikipedia.org/wiki/GraphPad_Software). GraphPad was developed by GraphPad Software, Inc.

Stata: This is a general-purpose statistical software package created in 1985 by StataCorp. Most of its users work in research, especially in the fields of economics, sociology, political science, biomedicine and epidemiology. Stata's capabilities include data management, statistical analysis, graphics, simulations, regression, and custom programming. It also has a system to disseminate user-written programs that lets it grow continuously (<https://en.wikipedia.org/wiki/Stata>).

3. IMPLEMENTATION AND DISCUSSION OF RESULTS

Implementation

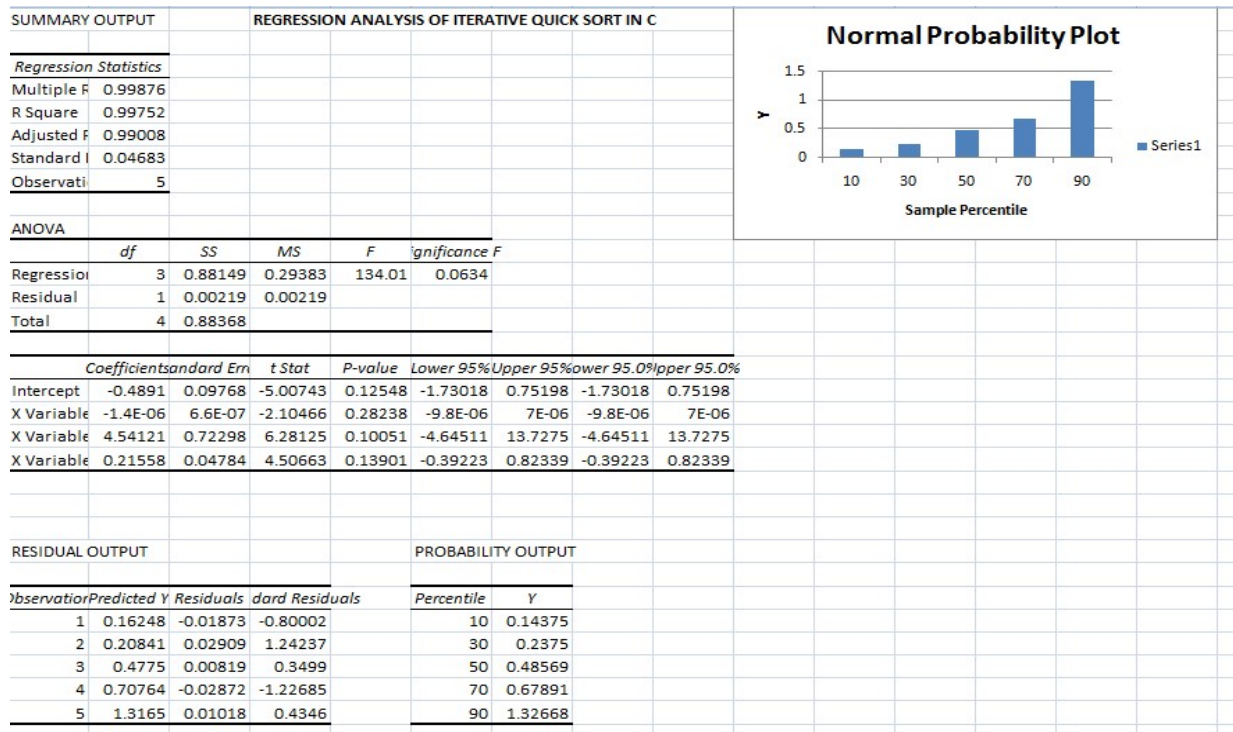
Based on the conclusion from the experiments conducted in Ayodele & Oluwade (2019), varying the parameters values (Data size, algorithm implementation style and programming language) impacts the energy consumption with different evolution patterns (see table 4.1).

Let E = Energy, x_1 = Data Size, x_2 = Execution Time, x_3 = Power
we have the following observations:

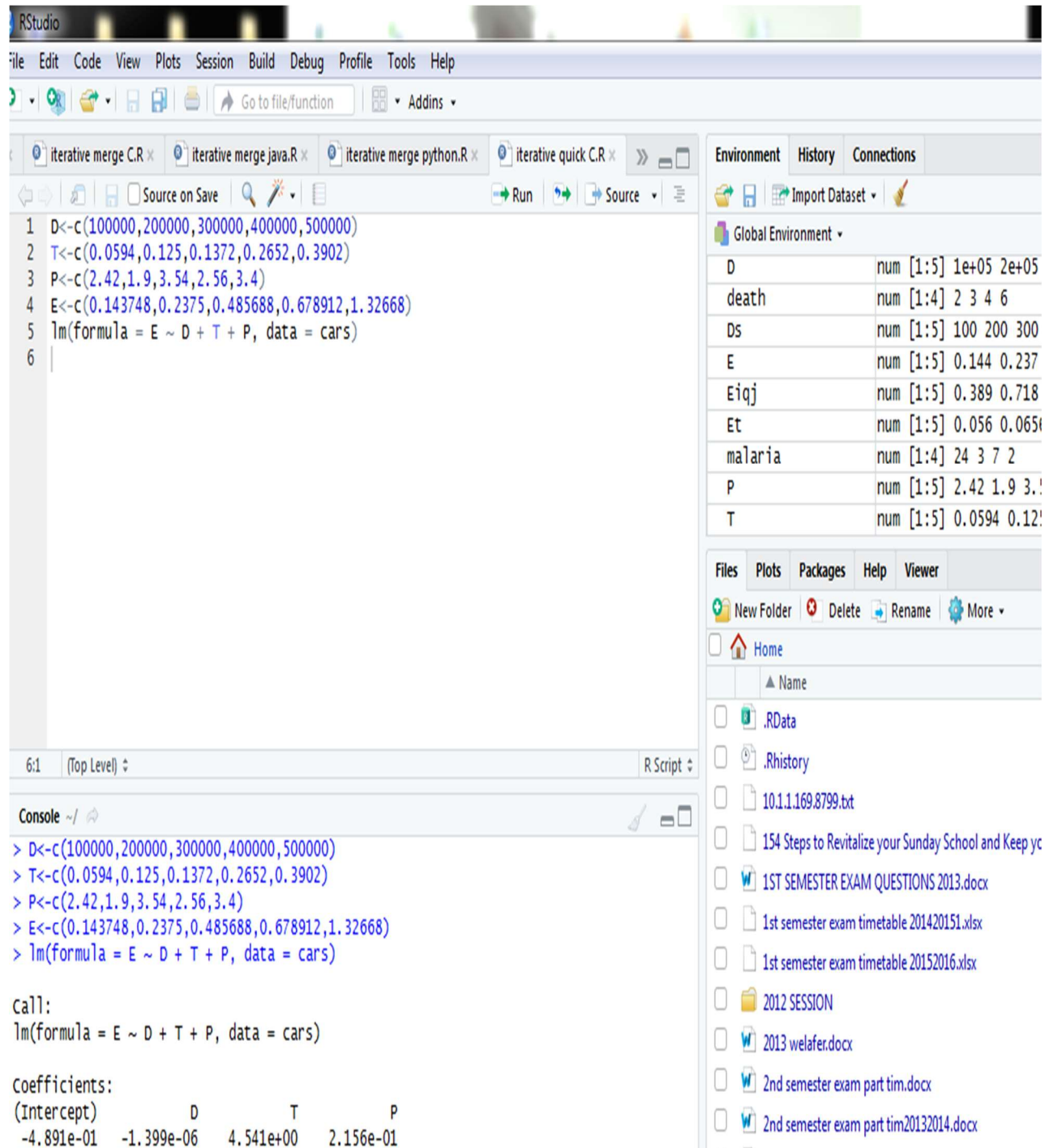
Two varying quantities are said to be in a relation of proportionality, if when they are multiplicatively connected to a constant, i.e, when either their ratio or their products yield a constant. The value of the constant is called the coefficient of proportionality. As Data Size increases, the Energy also increases, $E \propto x_i, i = 1(1)3$.

$$\Rightarrow E = \beta_0 + \sum_{i=1}^3 \beta_i x_i, \beta_i \in \mathbb{R} \quad (4.1)$$

To get the coefficients $\beta_0, \beta_1, \beta_2$ and β_3 , the regression analysis implementation using Microsoft Excel, SPSS, GraphPad, R, and Stata Software were used. Therefore, the regression models for predicting the energy consumption (Energy Efficiency) were developed as a function of Data Size, Execution Time and Power using five (5) statistical packages (Microsoft Excel, SPSS, GraphPad, R, and Stata). Predictive Regression Energy Model for Quick Sort Using Data Size, Execution time and Power in Microsoft Excel, SPSS, GraphPad, R, and Stata Software



USING R



The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for data creation and linear model fitting.


```

1 D<-c(100000,200000,300000,400000,500000)
2 T<-c(0.0594,0.125,0.1372,0.2652,0.3902)
3 P<-c(2.42,1.9,3.54,2.56,3.4)
4 E<-c(0.143748,0.2375,0.485688,0.678912,1.32668)
5 lm(formula = E ~ D + T + P, data = cars)
6
      
```
- Environment:** Lists objects in the Global Environment.

Object	Class	Value
D	num [1:5]	1e+05 2e+05
death	num [1:4]	2 3 4 6
Ds	num [1:5]	100 200 300
E	num [1:5]	0.144 0.237
Eiqj	num [1:5]	0.389 0.718
Et	num [1:5]	0.056 0.065
malaria	num [1:4]	24 3 7 2
P	num [1:5]	2.42 1.9 3.54
T	num [1:5]	0.0594 0.125
- Files:** Shows a list of files in the current directory, including .RData, .Rhistory, and various exam question documents.
- Console:** Displays the output of the R code execution.


```

> D<-c(100000,200000,300000,400000,500000)
> T<-c(0.0594,0.125,0.1372,0.2652,0.3902)
> P<-c(2.42,1.9,3.54,2.56,3.4)
> E<-c(0.143748,0.2375,0.485688,0.678912,1.32668)
> lm(formula = E ~ D + T + P, data = cars)

call:
lm(formula = E ~ D + T + P, data = cars)

Coefficients:
(Intercept)          D          T          P
-4.891e-01 -1.399e-06  4.541e+00  2.156e-01
      
```


Using GraphPad

GraphPad InStat - [DATASET1.ISD]

File Edit Data Steps Window Help

7/30/2019 11:08 PM

Multiple Regression Results

What equation fits the data the best?

$$[A:] = -0.4891 - 1.398E-06*[B:] + 4.541*[C:] + 0.2156*[D:]$$

Variable	Coefficient	SE	95% Confidence Interval
(constant)	-0.4891	0.09768	-1.730 to 0.7520
B:	-1.398E-06	6.646E-07	-9.843E-06 to 7.045E-06
C:	4.541	0.7230	-4.645 to 13.727
D:	0.2156	0.04784	-0.3922 to 0.8234

How good is the fit?

R squared = 99.75%.

This is the percent of the variance in A: explained by the model.

The P value is 0.0634, considered not quite significant.

The P value answers this question:

If there were no linear relationship among the variables, what is the chance that R squared would be that high (or higher) by chance?

Since P is high, the rest of the results will be of little interest.

Sum-of-squares	0.002193
SD of residuals	0.04683
R squared	0.9975
Adjusted R squared	0.9901
Multiple R	0.9988

Checklist ? What's next? Steps: 1st

Using Stata

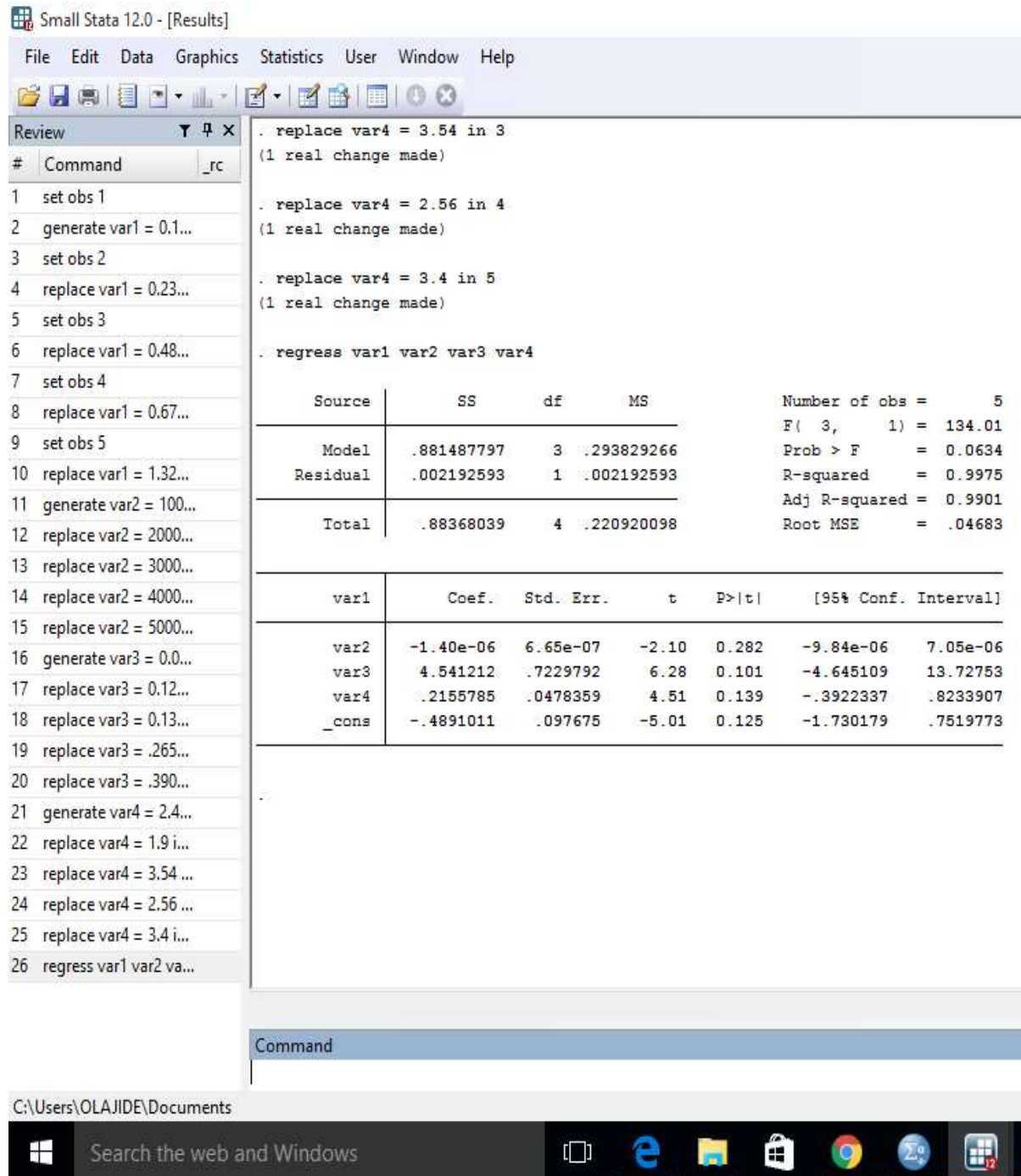


Figure 4.4 Regression Analysis of Iterative QuickSort Implementation in C Using Stata

Using SPSS

*Output1 [Document1] - IBM SPSS Statistics Viewer

File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help

Log
Regression
Title
Notes
Active Dataset
Variables Entered/Removed
Model Summary
ANOVA
Coefficients

a. All requested variables entered.
b. Dependent Variable: VAR00001

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.999 ^a	.998	.990	.0468251582	.998	134.010	3	1	.063

a. Predictors: (Constant), VAR00004, VAR00003, VAR00002

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.881	3	.294	134.010	.063 ^a
	Residual	.002	1	.002		
	Total	.884	4			

a. Predictors: (Constant), VAR00004, VAR00003, VAR00002
b. Dependent Variable: VAR00001

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.489	.098		-5.007	.125
	VAR00002	-1.399E-6	.000	-.471	-2.105	.282
	VAR00003	4.541	.723	1.275	6.281	.101
	VAR00004	.216	.048	.317	4.507	.139

a. Dependent Variable: VAR00001

Open output document

IBM SPSS Statistics Processor is ready

11:16 PM
7/31/2019

Table 4.1: Comparison of the Output of Regression Analysis of five Statistical Packages.

Package/Coefficient	β_0	β_1	β_2	β_3	ENERGY MODEL
MICROSOFT EXCEL	-0.4891	-0.0014e-03	+4.541213	+0.215578	$E_{iqc} = -0.4891 - 0.0014e-03Ds + 4.541213Et + 0.215578P$
R PACKAGE	-0.4891	-1.399e-06	+4.541	+0.2156	$E_{iqc} = -0.4891 - 1.399e-06Ds + 4.541Et + 0.2156P$
GRAPHPAD	-0.4891	-1.398E-06	+4.541	+0.2156	$E_{iqc} = -0.4891 - 1.399e-06Ds + 4.541Et + 0.2156P$
SPSS	-0.489	-1.399E-06	+4.541	+0.216	$E_{iqc} = -0.489 - 1.399e-06Ds + 4.541Et + 0.216P$
STATA	-0.4891011	-1.40E-06	+4.541212	+0.2155785	$E_{iqc} = -0.4891011 - 1.40e-06Ds + 4.541212Et + 0.2155785P$

Table 4.2: Comparison of the characteristics features

S/N	MICROSOFT EXCEL	R PACKAGE	GRAPHPAD	SPSS	STATA
1	User Friendly	Not user friendly	User friendly	User friendly	User friendly
2	Cost of software package Cheap	Free	Cost of software package is high	Cost of software package is high	Free trial version. Cost of software package is high
3	The users' interface is fair and gives a detailed result compared to others	The users' interface is poor compared to others statistical packages under observation	The users' interface is fair and gives a detailed result compared to others	The users' interface is fair and gives a detailed result compared to others	The users' interface is poor compared to others statistical packages under observation
4.	Memory usage for its installation is very low	Memory usage for its installation is very low	Memory usage for its installation is low	Occupy much space compared to others.	Memory usage for its installation is low
5	Knowledge of Programming not required	Knowledge of Programming required	Knowledge of Programming not required. Easy to understand and work with	Knowledge of Programming not required	Knowledge of Programming not required
6	Good and recommended for all beginners	Not Good and not recommended for the beginners	Good and recommended for the beginners	Not Good and not recommended for the beginners	Not Good and not recommended for the beginners
7	Result interpretation is easy	Result interpretation is not easy and it requires good expertise knowledge	Result interpretation is easy	Result interpretation is not easy and it requires good expertise knowledge	Result interpretation is not easy and it requires good expertise knowledge

4. DISCUSSION OF RESULTS

The output of regression analysis using SPSS is approximated irrespective of the number of decimal places specified during data analysis. The results generated by R package and GraphPad are the same. The result from Microsoft Excel is better than R package and GraphPad and has the advantage of user friendly environment. However, with Stata, the results have an extended approximate values, making its output to be reliable for solutions that requires high accuracy.

5. CONCLUSION

Statistical packages vary greatly with respect to accuracy, speed, reliability and reported results may be dependent on the specific package and version. Therefore one of the factors to be considered while choosing the statistical software to be used for data analysis is the intended use of the result of the regression analysis. However, Stata does the work with high accuracy in good speed and with the fair users' interface.

REFERENCES

1. Ayodele, O. S., & Oluwade, B. (2019). A comparative Analysis of Quick, Merge and Insertion Sort Algorithms using three programming Languages I: Execution Time. *African Journal of Mgt Information System* , 1-18.
2. Cavaliere, R. (2015). How to choose the right statistical software?—a method increasing the post-purchase satisfaction. *how to choose the right statistical software?- a method increasing the post- purchase satisfaction* , 585-598.
3. Michael Lewis-Beck, E. P. (2004). *The SAGE Encyclopedia of Social Science Research Methods*, Volume 1. Seattle, Washington: Amazon.com.
4. <https://en.wikipedia.org/wiki/Stata>
5. Godsey (2019). <https://towardsdatascience.com/how-to-choose-statistical-software-tools>, Accessed on July 1st, 2019.