Advances in Mathematical & Computational Sciences Journal,
Vol.7 No. 3, September 2019

Article Citation Format
Ayodele, O.S., Alabi, T.J & Ajayi, O.F. (2019): Comparative Analysis of Five Statistical Packages' Features and Output.
Journal of Advances in Mathematical & Computational Sciences
Vol. 7, No. 2. Pp 9-18

Article Progression Time Stamps

Article Type: Research Article
Manuscript Received 9th August, 2019
Final Acceptance: 12th September, 2019
Article DOI: dx.doi.org/10.22624/AIMS/MATHS/V7N3P2

Comparative Analysis of Five Statistical Packages' Features and Output

Ayodele, O.S., Alabi, T.J & Ajayi, O.F.
Kogi State Polytechnic
Lokoja, Nigeria
Correspondence Email: 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, Stata is highly recommended.

Keywords: Data manipulation, visualization, Data exploration and statistical analysis

1. INTRODUCTION

Statistics is the science of making effective use of numerical data relating to groups of individuals or experiments. It deals with all aspects of this, including not only the collection, analysis and interpretation of such data, but also the planning of the collection of data, in terms of the design of surveys and experiments. Traditional or manual method takes lot of time whereas statistical software makes the statistical analysis easier with better accurate result. Statistical software are specialized computer programs for analysis in statistics and econometrics. 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 S**ciences (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 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/GrapgPad_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, $E \propto x_i, i = 1(1)3.$

$$\Longrightarrow E = \beta_0 + \sum_{i=1}^{3} \beta_i x_i \, , \beta_i \epsilon \mathbb{R} \qquad (1)$$

To get the coefficients $\beta_0, \beta_1 \, , \beta_2 \, \text{and} \beta_3$, 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

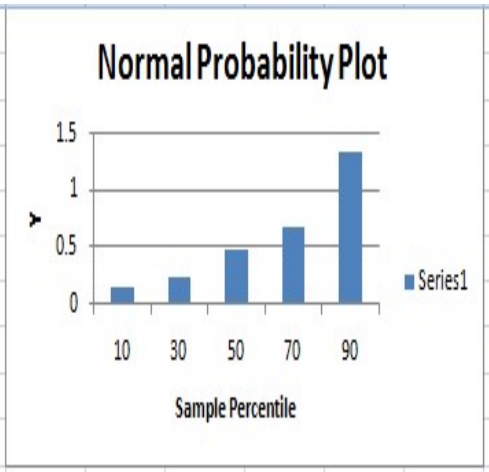
SUMMARY OUTPUT — REGRESSION ANALYSIS OF ITERATIVE QUICK SORT IN C

Regression Statistics	
Multiple F	0.99876
R Square	0.99752
Adjusted F	0.99008
Standard I	0.04683
Observati	5

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressio	3	0.88149	0.29383	134.01	0.0634
Residual	1	0.00219	0.00219		
Total	4	0.88368			

	Coefficients	*tandard Er*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-0.4891	0.09768	-5.00743	0.12548	-1.73018	0.75198	-1.73018	0.75198
X Variable	-1.4E-06	6.6E-07	-2.10466	0.28238	-9.8E-06	7E-06	-9.8E-06	7E-06
X Variable	4.54121	0.72298	6.28125	0.10051	-4.64511	13.7275	-4.64511	13.7275
X Variable	0.21558	0.04784	4.50663	0.13901	-0.39223	0.82339	-0.39223	0.82339



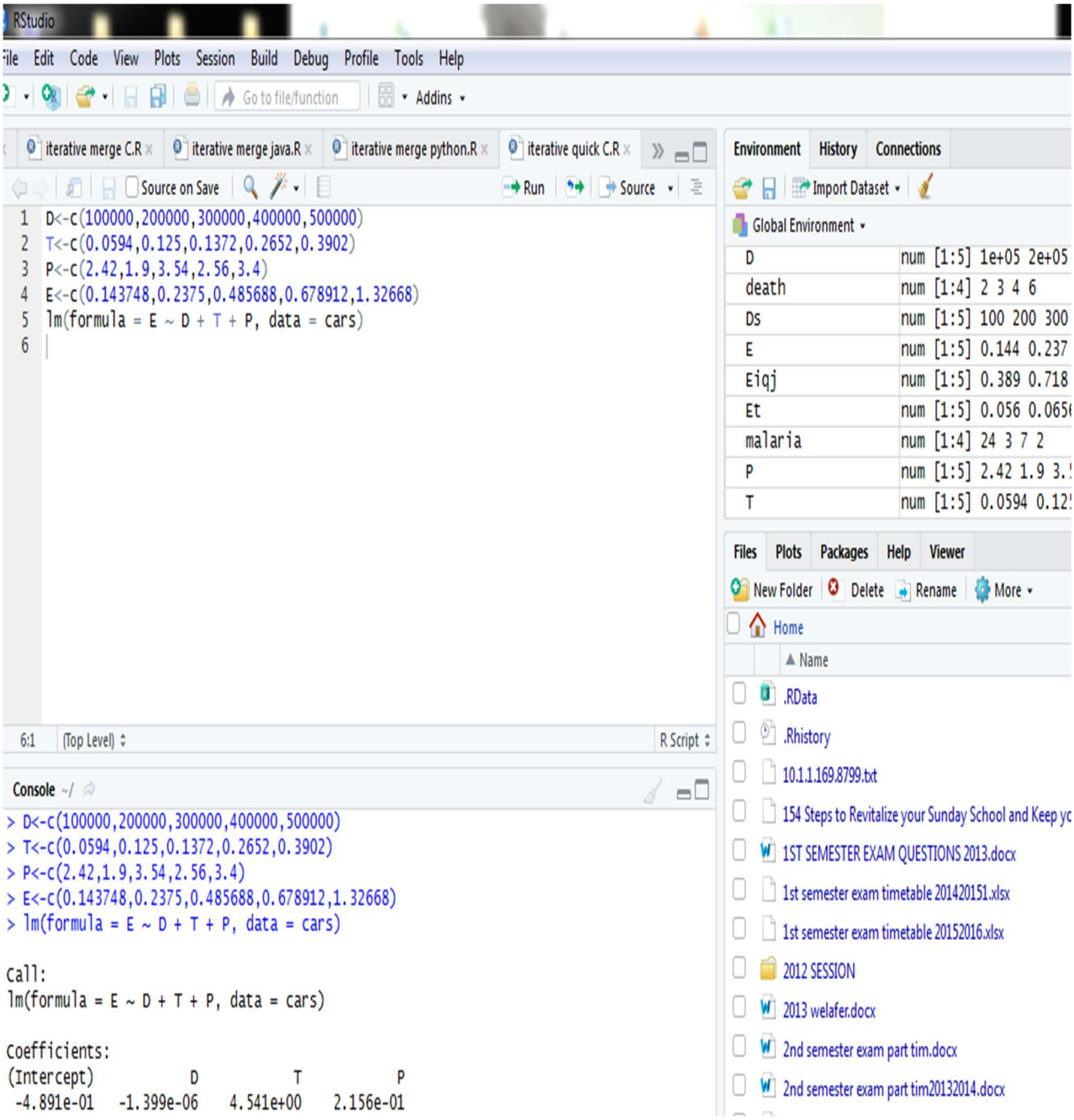
RESIDUAL OUTPUT

)bservatio	*Predicted Y*	*Residuals*	*dard Residuals*
1	0.16248	-0.01873	-0.80002
2	0.20841	0.02909	1.24237
3	0.4775	0.00819	0.3499
4	0.70764	-0.02872	-1.22685
5	1.3165	0.01018	0.4346

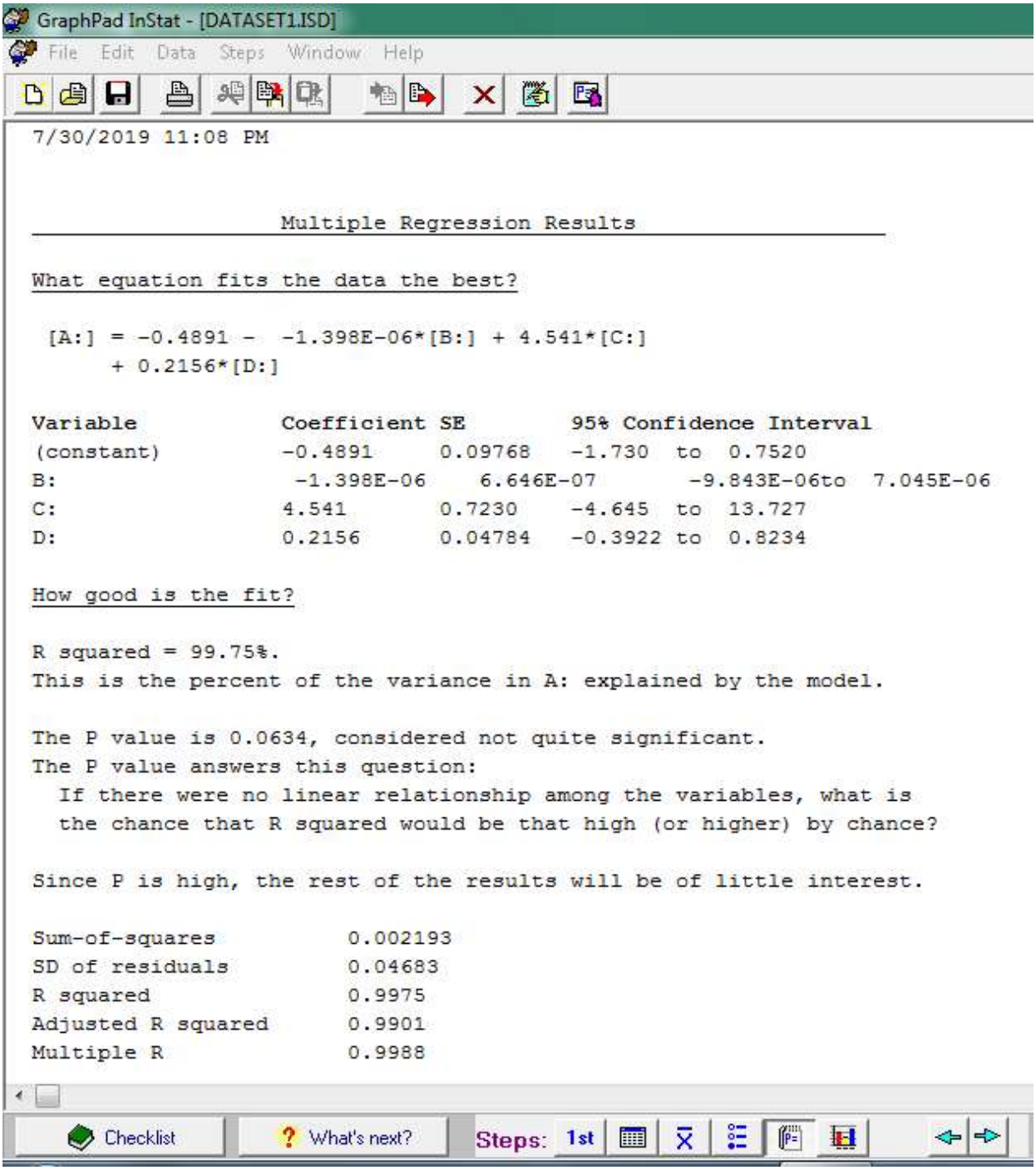
PROBABILITY OUTPUT

Percentile	*Y*
10	0.14375
30	0.2375
50	0.48569
70	0.67891
90	1.32668

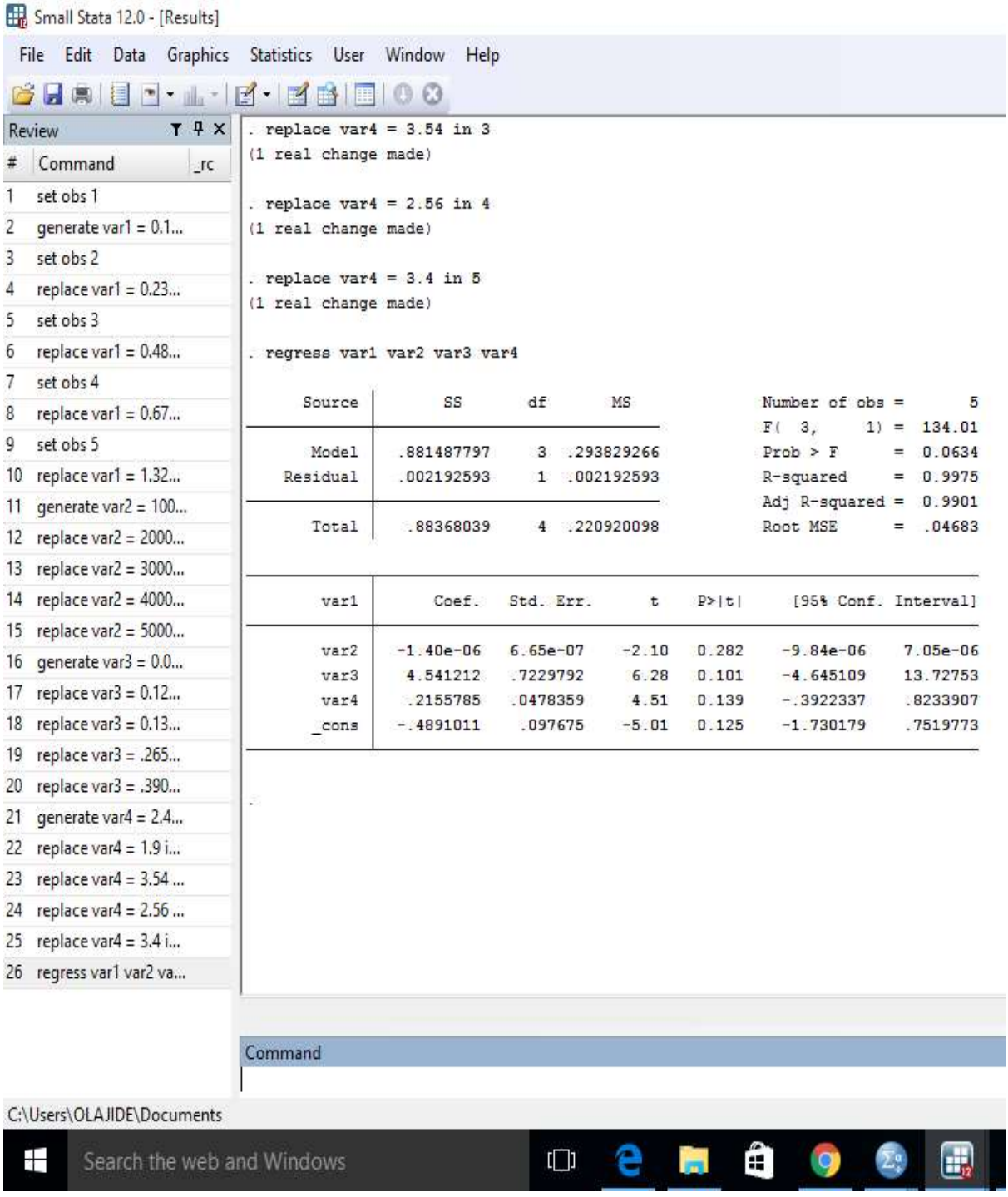
USING R


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
iterative merge C.R
iterative merge java.R
iterative merge python.R
iterative quick C.R
Source on Save
Run
Source
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
6:1 (Top Level)
R Script
Console ~/
> 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
Environment History Connections
Import Dataset
Global Environment
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.
T num [1:5] 0.0594 0.12
Files Plots Packages Help Viewer
New Folder Delete Rename More
Home
Name
.RData
.Rhistory
10.1.1.169.8799.txt
154 Steps to Revitalize your Sunday School and Keep yc
1ST SEMESTER EXAM QUESTIONS 2013.docx
1st semester exam timetable 201420151.xlsx
1st semester exam timetable 20152016.xlsx
2012 SESSION
2013 welafer.docx
2nd semester exam part tim.docx
2nd semester exam part tim20132014.docx

Using GraphPad



Using Stata


Small Stata 12.0 - [Results]
File Edit Data Graphics Statistics User Window Help
Review
Command _rc
1 set obs 1
2 generate var1 = 0.1...
3 set obs 2
4 replace var1 = 0.23...
5 set obs 3
6 replace var1 = 0.48...
7 set obs 4
8 replace var1 = 0.67...
9 set obs 5
10 replace var1 = 1.32...
11 generate var2 = 100...
12 replace var2 = 2000...
13 replace var2 = 3000...
14 replace var2 = 4000...
15 replace var2 = 5000...
16 generate var3 = 0.0...
17 replace var3 = 0.12...
18 replace var3 = 0.13...
19 replace var3 = .265...
20 replace var3 = .390...
21 generate var4 = 2.4...
22 replace var4 = 1.9 i...
23 replace var4 = 3.54 ...
24 replace var4 = 2.56 ...
25 replace var4 = 3.4 i...
26 regress var1 var2 va...
. replace var4 = 3.54 in 3
(1 real change made)
. replace var4 = 2.56 in 4
(1 real change made)
. replace var4 = 3.4 in 5
(1 real change made)
. regress var1 var2 var3 var4
Source | SS df MS
Model | .881487797 3 .293829266
Residual | .002192593 1 .002192593
Total | .88368039 4 .220920098
Number of obs = 5
F(3, 1) = 134.01
Prob > F = 0.0634
R-squared = 0.9975
Adj R-squared = 0.9901
Root MSE = .04683
var1 | Coef. Std. Err. t P>|t| [95% Conf. Interval]
var2 | -1.40e-06 6.65e-07 -2.10 0.282 -9.84e-06 7.05e-06
var3 | 4.541212 .7229792 6.28 0.101 -4.645109 13.72753
var4 | .2155785 .0478359 4.51 0.139 -.3922337 .8233907
_cons | -.4891011 .097675 -5.01 0.125 -1.730179 .7519773
Command
C:\Users\OLAJIDE\Documents
Search the web and Windows

Using SPSS

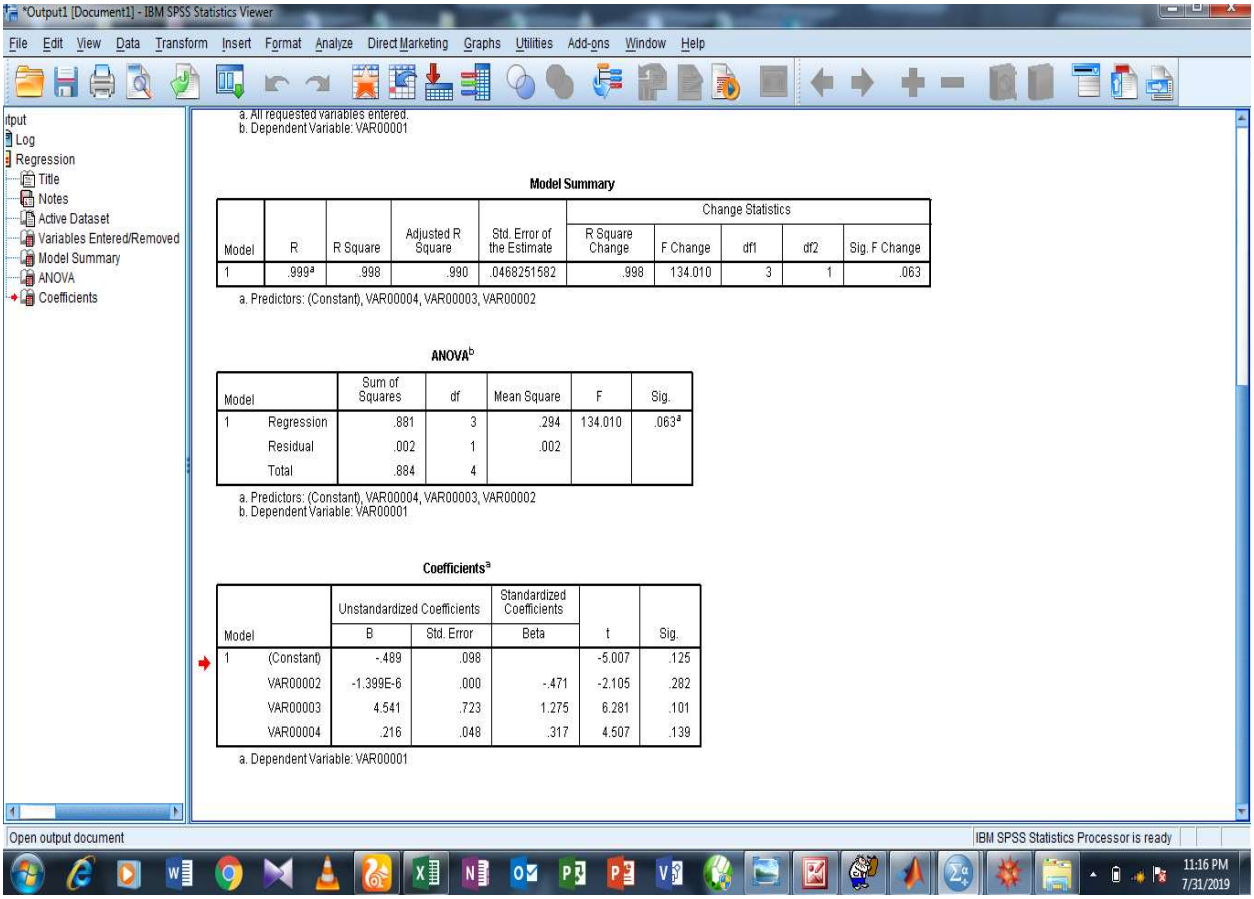


Table 1: Comparison of the Output of Regression Analysis of five Statistical Packages.

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}=-4891011–1.40e-06Ds +4.541212Et +0.2155785P

Table 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 is 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 incresing 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.