SUDAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

College of Graduate Studies

Big Data Analysis using Weka Machine Learning Program and SPSS Package: A Comparative Study

تحليل البيانات الكبيره باستخدام WEKA لغه تعلم الاله وحزمه SPSS:

دراسة مقارنه

A thesis Submitted in filfittment of requirement for the degree of PhD in statistics

by:

Rawia Hashem Mustafa

Supervisor: prof. Amin Ibrahim Adam

Co- Supervisor: Dr. Afra Hashim Abdallateef

July 2020

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

قال تعالى:﴿ اقْرَأْ بِاسْمِ رَبِّكَ الَّذِي خَلَقَ (1) خَلَقَ الْإِنْسَانَ مِنْ عَلَقٍ (2) اقْرَأْ وَرَبُّكَ الْأَكْرَمُ (3) الَّذِي عَلَّمَ بِالْقَلَمِ (4) عَلَّمَ الْإِنْسَانَ مَا لَمْ يَعْلَمَ(5) ﴾

صدق الله العظيم

سورة العلق الآيات من 1-5

Dedication

To my mother and father

To brothers and sisters

To my husband and daughters who offered great advise during my study

Abstract

Companies are currently rich in huge amounts of data but are weak in information extracted from data. This massive data is a valuable resource. Although the concept of big data is still new, many international companies are relying on it to make strategic decisions.

This study examined the regression analysis, which is part of Multivariate Analysis, which was an implemented step method on real data and the efficiency of the regression analysis when increasing the size of the sample tested, The importance of this study is to compare the method of statistical analysis, keep in mind the optimal use of available resources in computers such as Random-Access memory (RAM) and processor speed to reach the results to be the best in terms of time spent on analysis after confirmation of the validity of the steps involved in getting the prediction.

The study used the descriptive statistical approach to describe the study variables, and the analytical statistical approach to obtain the study results using the SPSS20, WEKA programs.

The aim of the study was to make a comparison between the WEKA program indicators and the SPSS package to determine which of them is efficient in large samples.

The study concluded that using WEKA software because gives better results than using SPSS program.

And Increasing the sample size increases the efficiency of the WEKA program indicators compared with the SPSS program indicators when large samples.

The study recommends to using the weka program in the case of large samples because it is more efficiency in prediction parameters compared with the SPSS program indicators in the same parameters.

المستخلص

الشركات الكبيرة في الوقت الحالي غنية بكميات هائلة من البيانات ولكنها ضعيفة في المعلومات المستخرجة من البيانات .وتعتبر هذه البيانات الضخمة مصدر قيما للمعلومات، ورغم أن مفهوم البيانات الكبيرة لا يزال جديدا، فإن العديد من الشركات الدولية أصبحت تعتمد عليها لاتخاذ القرارات الاستراتيجية.

تناولت هذه الدراسة تحليل اداه WEKA والحزمة SPSS في regression Analysis كواحدة من طرق المستخدمة في التنبؤ الذي ينحدر من تحليل متعدد المتغيرات (Multivariate Analysis) , حيث تم تطبيق خطوات عمل الطريقة علي بيانات حقيقية و اختبار مدي كفاءة الطريقة في تحليل الانحدار عند زيادة حجم العينة.

تتمثل اهمية هذه الدراسة في انها تهتم بمقارنة طريقة احصائية في التحليل مع مراعات الاستخدام الامثل للموارد المتاحة في الحواسيب (Personal computers) مثل الذاكرة المؤقتة (RAM) وسرعة المعالج (Processor) للوصل الي نتائج تكون هي الافضل من حيث الزمن المستغرق في التحليل بعد التأكد من صحة الخطوات المتبعة في الحصول علي التنبؤ.

وأتبعت الدراسة المنهج الاحصائي الوصفي لوصف متغيرات الدراسة , والمنهج الاحصائي التحليلي للوصول الي نتائج الدراسة باستخدام البرامج SPSS20, WEKA.

وكان لهدف من الدراسة عمل مقارنه بين مؤشرات برنامج WEKA وحزمه SPSS لتحديد ايهما اكفاء في العينات الكبيرة.

وتوصلت الدراسة الي ان استخدام برنامج WEKA يعطي نتائج أفضل من استخدام حزمه البيانات SPSS.

زيادة حجم العينة تزيد من كفاءة مقدرات برنامج WEKA عند زياده حجم العينه مقارنه مع مؤشرات برنامج SPSS عند العينات الكبيره.

وتوصي الدراسة باستخدام weka نسبه لفعاليتها في تقدير المعلمات مقارنه بتقدير نفس المعلمات في spss.

Table of Contents

List of Tables

List of figures

Preface

Statistics is an important part in big data because many statistical methods can use for big data analysis. The aim of statistics is to estimate population using the sample extracted from the population, so statistics is to analyze data from the sample.in big data environment, we can get the big dataset closed to the population by the advanced program such as WEKA machine learning and SPSS package.

We can analyze entire part of big data like the population of statistics. But it may be impossible to analyze the entire data because it's huge data volume, in this research we concerned with method of finding regression equation Predictive modeling is one of popular mathematical techniques having in common the goal off finding a mathematical relationship between a target, response, or "dependent" variable and various predictor or "independent" variables with the goal in mind of measuring future values of those predict.

Regression analysis establishes a relationship between a dependent or outcome variable and a set of predictors.

1.2 Problem of the study

With the significant development of constantly updated data and information, whether directly accessible through simple search or indirectly through advanced search with different search engines, there are databases and engine bases whose data can only be access through exploration. different companies and researchers in the field of data search have sought to produce tools that help extract information from data in different databases, including what is available free or paid, this study tries to highlight the most important tools of analysis big data program of an open source data program and their evaluation to determine its advantages,

The main problem is finding an equation to predict the number of comments on the social networking site Facebook in the WEKA program to show future values as a simple example of big data.

1.3 Importance of the study

The practical importance is the lack of studies on the analysis of large data, especially open source tools, and the study of what WEKA is as a program of learning machine to analyze data and its functions and development.

Then evaluate the slope equation using the WEKA tool when the large samples are compare with the regression equation when using SPSS.

The study seeks to achieve many objectives such as analyzing large data through statistical package SPSS and studying what is the data analysis program WEKA program and its functions and applications.

1.4 Objective of the study

1. General objective to Measure the relative efficiency of the WEKA program by focusing on the RMSE and RMAE and estimating when the sample size increases each time.
2. Specific objective to resolve the ongoing controversy in the problem of analysis of large data and the establishment of linear regression model to predict future values and by achieving the sub-goals Adopting methods of scientific methodology to construct a linear regression model, improve the results of linear regression model.
3. Study the efficiency of the program to analyze regression by measuring its effect on linear regression model indicators.
4. Measure the relative efficiency of the WEKA program by focusing on the MSE and MAE and estimating when increasing the sample size each time
5. To Study between measure of efficiency and measure of accuracy such as analyzing big data through statistical package

SPSS and WEKA program to determine which better in big data.

1.5 Source of the study

The study based on Data processed for scientific study purposes were use from the machine and intelligent systems education center site UCLA, which is a public study university located in Irvine, California, USA, and is one of 10 universities in the California University System. The website provide real datasets ready for scientific studies.

The data is link to the volume of comments about a publication and the study data revolve around expected comments in a particular publication in order to predict the size of expected comments. The data relates to the number of (199030) comments.

1.6 Hypotheses of the study

1. Main hypotheses, no significant difference between simple regression equation in big data using WEKA program and SPSS package.
2. There is no a statistically significant difference between multiple regression equation in big data using WEKA program and SPSS package.
3. The correlation between big data variables is significant.

4. WEKA the more efficient for estimation big data.

1.7 Structure of research:

Chapter 1: introduction

Chapter 2: literature review

Chapter 3: statistical Method for analyzing big data

Chapter 4: Results and Recommendations

1.8 Methodology of the study

In this study, a descriptive and analytical method will be use. The tools that will be used in the implementation of this research are statistical and computer tools. In general, the approach of this research is a pilot approach, which depends primarily on the practical application and to ensure the above objectives, the researcher used WEKA and SPSS program as analytical tools to estimate and analyze data:

- To resolve the ongoing controversy in the problem of analysis of big data and the establishment of linear regression model to predict future values and by achieving the sub-goals:
- Adopting methods of scientific methodology to construct a linear regression model
- Make sure that the linear regression model data is of the same value.

- Improve the results of linear regression model and sure the reliability

1.9 previous studies

The related works to our research are:

Buza Krisztian has written “Feedback Prediction for Blogs"

Buza develop an industrial proof-of-concept demonstrating the automatic analysis of the documents on Hungarian blogs. The author has trained various regression models by considering various features of the blogs and measured the results by using the parameter AUC (Area under curve). The result shows that the regression models outperform than naive models. [1]

M.Tsagkias, W. Weerkamp and M. de Rijke "Predicting the Volume of Comments on Online News Stories”.[2]

Even, classifiers can be use to categorize the comment volumes in specific classes like what They report on "Predicting the Volume of Comments on Online News Stories” prediction the comment volume of news before their publication using Random Forest classifier based on the set of five features, i.e. surface, cumulative, textual, semantic, and real-world features. It addresses the task in two steps; first binary

[1]Buza Krisztian, “Feedback Prediction for Blogs”, Springer International Publishing on Data Analysis, Machine Learning and Knowledge

[2]M.Tsagkias, W. Weerkamp, M. de Rijke, “Predicting the Volume of Comments on Online News Stories

classification of articles with the potential to receive comments, and second to classify articles with "low volume" and "high volume". Outcomes show better results for binary classification and evaluated that the textual and semantic features are strong performers among others.

Balali, A. and Rajabi, A. and Ghassemi, S. andAsadpour, M. and Faili, H., "Content diffusion prediction in social networks"[3]

In this paperhave made analysis on the content and publish time of online news agencies, to detect effective factors of diffusing contents in public. It has also used the Random Forest Classifier to classify articles in three categories, i.e. without commented, moderately commented (1-6) and highly commented (>6).The proposed model has made predictions with more than 70% accuracy and reports that the publish date and a weight introduced for content measure, were most informative features. The results can be refine by considering important days (i.e. elections, festivals, and holidays) and geographical features in prediction.

[3]Balali, A. and Rajabi, A. and Ghassemi, S. andAsadpour, M. and Faili, H., "Content diffusion prediction in social networks", Information and Knowledge Technology (IKT), 5th IEEE Conference, 2013, pp. 467-471. doi: 10.1109/IKT.2013.6620114.

M. Tsagkias, W. Weerkamp, M. de Rijke, "News Comments: Exploring, Modeling, and Online Prediction" proceedings of the 32nd European conference.

They have shown the dynamics of user generated comments on seven different news websites, using the log-normal and the negative binomial distributions and predicted the comment volume using Linear model and enable comparison across various news sites. The results showed that prediction of long-term comments volume is possible with small error after 10 source-hours observations.[4]

Jamali, S. and Rangwala, H. have worked on social bookmarking website Digg.com. They published the paper "Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis"

They have used comment information; it defined a co-participation network between users and studied the behavioral characteristics of users. It measured the entropy and inferred that the users at Digg are interested in wide range of topics. Using a classification and regression framework, it has predicted the popularity of online content based on comment data and social network derived features. It reported a one to four percent loss in classification accuracy while predicting the popularity metric by using only first few hours of comment data as compare to all the available

[4]M. Tsagkias, W. Weerkamp, M. de Rijke, "News Comments: Exploring, Modeling, and Online Prediction" proceedings of the 32nd European conference

comment data. The results can further be improve by analyzing the polarity of the comments.[5]

Various topic models can used to extract the hidden topics in post's content.

In 4th International AAAI Conference on Weblogs and Social MediaYano, Tae, and Noah A. Smith. Have written "What's worthy ofComment? Content and Comment Volume in Political Blogs". Theyhave worked on political blogs using Latent Variable topic model and analyzed the relationship between the content and comment volume. It has also used Naïve Bayes model for binary prediction task i.e. high volume or low volume and evaluated the prediction using precision, recall and F1 measures. It concluded that the modeling topics can improve recall while predicting high volume posts.[6]

Negi, S. and Chaudhury "Predicting User-to-content Links in Flickr Groups"[7]

Has predicted the formation of user-to-content links in Flickr Groups to predict the chance that a user will comment or like an image updated by another user. It has considered both the community effect using Transactional Mixed Membership Stochastic Block (TMMSB) model and content effect using Latent Dirichlet Allocation (LDA) for predicting

---

[5]Jamali, S. and Rangwala, H., "Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis",

[6]Yano, Tae, and Noah A. Smith. "What's Worthy of Comment? Content and Comment Volume in Political Blogs"

[7]Negi, S. and Chaudhury "Predicting User-to-content Links in Flickr Groups"

user-to-content links. The time zone effects can used in future in order to make results more accurate. Summarization of the user's comments is even more difficult task as these are usually mix with different opinions, specifically in case of restaurants where different opinions refer to different dishes but evaluated as an overall score of restaurants.

Rong Zhang, Zhenjie Zhang, Xiaofeng He, Aoying Zhou, "Dish Comment Summarization Based on Bilateral Topic Analysis".[8]

has presented a new approach for comment summarization in the context of restaurants in there paper, They used the real-world comments, crawled from Yelp and Damping, the most popular English and Chinese restaurant review web sites. Using the attributes of the dishes and the user's remarks on the attributes as two independent dimensions in the latent space, it constructed a bilateral topic model that is combine with the opinionated word extraction and clustering-based selection algorithms; it provides a high-quality summary on the restaurants as well as the dishes served by the restaurants. This concept can further used for wider applications like for various selling goods or services.

[8]Rong Zhang, Zhenjie Zhang, Xiaofeng He, Aoying Zhou, "Dish Comment Summarization Based on Bilateral Topic Analysis".

In contrast to these works, we have focused on leading platform i.e. Facebook, a leading platform and targeted the regression models as linear Regression and multiple linear regression.

Chapter 2

Literature review

2.1 Preface

The digital data is produce as part of the use of devices connected to the Internet. Thus, smart phones, tablets and computers transmit data about their users. Connected smart objects convey information about consumer's use of everyday objects.

Apart from the connected devices, data come from a wide range of sources: demographic data, climate data, scientific and medical data, energy consumption data…etc. All these data provide information about the location of users of the devices, their travel, their interests, their consumption habits, their leisure activities, and their projects and so on. But also, information on how the infrastructure, machinery and apparatus are used. With the ever-increasing number of Internet and mobile phone users, the volume of digital data is growing rapidly. Today we are living in an informational society and we are moving towards a Knowledge Based Society. In order to extract better knowledge, we need a bigger amount of data. The Society of information is a society where information plays a major role in the economic, cultural and political stage.[9]

[9] H. Wimmer & L. M. Powell (2015). "*A Comparison of Open Source Tools for Data Science*". Proceedings of the Conference on Information Systems Applied Research Wilmington, North Carolina USA

2.2 Definition of big data

The term "Big Data" refers to the evolution and use of technologies that provide the right user at the right time with the right information from a mass of data that has been growing exponentially for a long time in our society. The challenge is not only to deal with rapidly increasing volumes of data but also to the difficulty of managing increasingly heterogeneous formats as well as increasingly complex and interconnected data.[10]

Being a complex polymorphic object, its definition varies according to the communities that are interested in it as a user or provider of services. Invented by the giants of the web, the big data presents itself as a solution designed to provide everyone a real-time access to giant databases. [11]

Big data is a very difficult concept to define precisely, since the very notion of big in terms of volume of data varies from one area to another. It is not define by a set of technologies; on the contrary, it defines a category of techniques and technologies. This is an emerging field, and as we seek to learn how to implement this new paradigm and harness the value, the definition is changing.

[10] Exploring, Modeling, and Online Prediction", ECIR'2010 Proceedings of the 32nd European conference on Advances in Information Retrieval, Springer

[11] Ishwarappa, and J. Anuradha, “A Brief Introduction On Big Data 5VsCharacteristics And Hadoop Technology”. Procedia Computer Science48, pp. 319-324, 2015.

2.3 Big data history

Big data is a long evolution of capturing and using of data and not a new phenomenon. Big data is the future act that will bring change in the way we run society, just like the other developments in storage of data, processing of data and internet. The ancient history of data is when humans used tally sticks for storing and analysis of data about C 18,000 BCE. The tribal peoples used to mark notches into bones or sticks for calculations, which would make them predict about how long their food would last. One of the earliest prehistoric data storage is Ishango Bone now known as Uganda, which was discover in 1960. Then in C 2400, BCE came the very first device particularly for performing calculations- Abacus.[12]

Our first libraries also appeared in this time period which represented our initial step towards mass storage. Then in the period of 300 BC-48 AD the library containing largest collection of data of the historic world which covered pretty much everything which we learned so far was destroyed by Romans accidentally.[13] Then started the early stage of

[12]Negi, S. and Chaudhury, S., "Predicting User-to-content Links in Flickr Groups", Advances in Social Networks Analysis and Mining (ASONAM), IEEE/ACM International Conference, 2012 pp. 124-131. doi: 10.1109/ASONAM.2012.31.

[13]Dr. K. J. Begum & Dr. A. Ahmed, "*The Importance of Statistical Tools in Research Work*". International Journal of Scientific and Innovative Mathematical Research (IJSIMR)

modern data storage[14]. In 1928, a German- Austrian engineer Fritz Plummer invented a magnetic tape, which stored information magnetically. Then came the Business Intelligence and start of large data centers where ideas of relational database and Material Requirement Planning systems were out forward.[15]

In 1989 the first use of the term big data was done by Erik Larson in the Harpers Magazine where he said that "The keepers of big data say they are doing it for the consumer's benefit. But data have a way of being used for purposes other originally intended". The birth of World Wide Web took place that kicked internet into gear in 1991. Google search engine debut in the year 1997. After a couple of years in 1999 big data term appeared in a research paper published by the association for computing Machinery. In that, storing large amounts of data and inadequate space for storage as well as analysis difficulties were highlight.

2.4 Behavioral types of big data

The data is been Categorized into many types according to behavior:

---

[14]Rahman, M.M. , "Intellectual knowledge extraction from online social data", Informatics, Electronics Vision (ICIEV), IEEE International Conference, 2012, pp. 205-210. doi:10.1109/ICIEV.2012.6317392

[15]Rong Zhang, Zhenjie Zhang, Xiaofeng He, Aoying Zhou, "Dish Comment Summarization Based on Bilateral Topic Analysis" Data Engineering (ICDE), 31st IEEE International Conference, 2015, pp. 483 – 494. doi: 10.1109/ICDE.2015.7113308

i. **Structured Data**

The data stored in relational databases table in the format of row and column. They have fixed structures and these structures are define by organizations by creating a model. The model allows to store, process as well as gives permission to operate the data. The model defines the characteristics of data including data type and some restriction on the data. Analysis and storing of structured data are very easy. Because of high cost, limited storage space and techniques used for processing, causes **Relational Database Management System** (RDBMS) the only path to store and process the data effectively. Programming language called **Structured Query Language** (SQL) is use for managing this type of data.

ii. **Unstructured data**

Data without any specific structure and due to this could not be stored in a row and column format is unstructured data. The data is contradictory to that of structured data. It cannot be stored in a databank. Volume of this data is growing extremely fast which is very tough to manage and analyze it completely. To analyze the unstructured data advanced technology knowledge is need.

iii. **Semi-structured data**

Data, which is in the form of structured data, but it does not fit the data model is semi-structured data. It cannot be stored in the form of data table, but it can be stored in some particular types of files, which hold some specific marker or tags. These markers are distinguish by some specific rule and the data is enforce to be stored with a ranking. This form of data increased rapidly after the introduction of the World Wide Web where various form of data needs medium for interchanging the information like XML and JSON[16]

2.5 **Big data analytics**

Big data analytics is a method to uncover the hidden designs in large data, to extract useful information that can be divide into two major sub-systems: data management and analysis.[17]

Big data analytics is a process of inspecting, differentiating and transforming big data with the goal of identifying useful information, suggesting conclusion and helping to take accurate decisions. Analytics include both data mining and communication or guide decision making. The analytics is concerned with the entire methodology[18]. Big data

[16]Buza Krisztian, "Feedback Prediction for Blogs", Springer International Publishing on Data Analysis, Machine Learning and Knowledge Discovery,2014

[17]Balali, A. and Rajabi, A. and Ghassemi, S. andAsadpour, M. and Faili, H., "Content diffusion prediction in social networks", Information and Knowledge Technology (IKT), 5th IEEE Conference

[18]Analytics Vidhya https://www.analyticsvidhya.com/blog/2014/03/sas-vs-vs-python-tool-learn/

analytics is use by all most all sectors for increasing productivity and revenue with decrease cost. It helps by optimizing funnel conversion, behavioral analytics, predictive support, market basket analysis, and pricing optimization, predicts security threat, fraud detection … etc.[19]

Big data analytics make sense of large volumes of data having variety of data in its raw form that lacks a data model.[20] Organizations collect, store and analyze massive amounts of data, which is referred as big data. Collecting and storing such huge amount of data has little value but analyzing gives tremendous value to the data. This analyzed data helps in decision-making and many other things.[21] Big data size range from few dozen terabyte to many peta bytes in a single data set. There are obvious difficulties like capturing data, storing, analyzing, visualizing, sharing... etc. even the data gained are not in a single format rather than they vary tremendously from structured, unstructured and even semi structured. There is a need for exert advanced analytical techniques on big data and this is where big data analytics helps. The analytics process is use to obtain previous unknown, useful hidden patterns, to extract useful

[19] Yano, Tae, and Noah A. Smith. "What's Worthy of Comment? Content and Comment Volume in Political Blogs"

[20] K. Rangra ,Dr. K. L. Bansal, "Comparative Study of Data Mining Tools", International Journal of Advanced Research in Computer Science and Software Engineering

[21] A.Komathi, T.Ramya, M. Shanmugapriya, V. Sarmila, "A Novel Comparative Study on Data Mining Tools

unknown relationships. Association rules, clustering, regression …etc. are the advanced analytical processes commonly used most[22].

[22] M.Tsagkias, W. Weerkamp, M. de Rijke, "Predicting the Volume of Comments on Online News Stories",CIKM"09 Proceedings of the 18th ACM conference on Information and knowledge management

2.6 Characteristics of Big Data

The term Big Data refers to gigantic larger datasets (volume); more diversified, including structured, semi-structured, and unstructured (variety) data, and arriving faster (velocity) than before. These are the 3V.

- **Volume:**

  Represents the amount of data generated, stored and operated within the system. The increase in volume is explained by the increase in the amount of data generated and stored, but also by the need to exploit it.

- **Variety:**

  Represents the multiplication of the types of data managed by an information system. This multiplication leads to a complexity of links and link types between these data. The variety also relates to the possible uses associated with a raw data.

- **Velocity:**

  Represents the frequency at which data is generated, captured, and shared. The data arrive by stream and must be analyzed in real time.

To this classical characterization, two other "V"s are important:

- **Veracity:** level of quality, accuracy and uncertainty of data and data sources.

- **Value:** the value and potential derived from data.

2.7 WHAT IS BIG DATA ANALYTIC?

Big Data generally refers to data that exceeds the typical storage, processing, and computing capacity of conventional databases and data analysis techniques. As a resource, Big Data requires tools and methods that can be apply to analyze and extract patterns from large-scale data.[23] The analysis of structured data evolves due to the variety and velocity of the data manipulated. Therefore, it is no longer enough to analyze data and produce reports, the wide variety of data means that the systems in place must be capable of assisting in the analysis of data. The analysis consists of automatically determining, within a variety of rapidly changing data, the correlations between the data in order to help in the exploitation of it.

Big Data Analytics refers to the process of collecting, organizing, analyzing large data sets to discover different patterns and other useful information. Big data analytics is a set of technologies and techniques that require new forms of integration to disclose large hidden values from large datasets that are different from the usual ones, more complex, and

[23] M.Tsagkias, W. Weerkamp, M. de Rijke, "Predicting the Volume of Comments on Online News Stories",CIKM"09 Proceedings of the 18th ACM conference on Information and knowledge

of a large enormous scale. It mainly focuses on solving new problems or old problems in better and effective ways.[24]

2.8 Levels of big data analytics

Big data analytics developing and implementation is not an easy task, especially when you do not have a data driven culture. Data driven culture is a pre-requisite for big data successful implementation. The right start to big data is to have an understanding of what is it and what can it do to the organization and from there proof of concept with multi-disciplinary team starts developing. This proof of concept is vital to the organization and for becoming data driven. There are 5 levels of big data maturity within an organization. First level: infancy phase- this is the phase where one starts understanding big data and develops proof of concepts.

Second level: Technical adoption: different big data technologies are implement. This will enable the organization to develop new proof of concepts faster and better. Third level: business adoption- more in deep analysis of structured and unstructured data which results in more sharp, accurate and better decision making of company. Fourth level: Enterprise adoption- the big data adoption across enterprise, which results in united

[24] Jamali, S. and Rangwala, H., “Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis

predictive insights of organization. At this level big data analytics has become an integral part of organization.

Fifth level: Data & Analytics as a service- at this level the organization operates as a “data service provider”. Organization has integrated big data analytics in all levels and now can be seen as ‘data companies’ no matter what product and service they provide.

2.9 Types of Big Data Analytics

After collection data we need to start analyzing it. There are types of analytics, which should use for different types of data. There are four types of analytics.

I. Descriptive Analytics

Descriptive analysis also known as data mining operates what is happening in real-time. It is one of the simplest types of analytics as it converts big data into small bytes. The result is monitor through e-mails or dashboard. It is use by majority of organizations. Descriptive analysis examines historical electricity usage to plan power need and set prices (Figure 1).[25]

It consists of asking the question: What is happening?

It is a preliminary stage of data processing that creates a set of historical data. Data mining methods organize data and help

[25] Dr. K. J. Begum & Dr. A. Ahmed, “*The Importance of Statistical Tools in Research Work*”. International Journal of Scientific and Innovative

uncover patterns that offer insight. Descriptive analytics provides future probabilities and trends and gives an idea about what might happen in the future.

II. Diagnostic Analytics

Diagnostic analysis looks to the past information and let us know how, what and why happened. It is usually use to uncover any hidden patterns, which help for complete root cause as well as identify any factors that are directly or indirectly causing effect. Diagnostic analysis is majorly use in social media for analyzing the number of posts, shares… etc.

It consists of asking the question: Why did it happen?

Diagnostic analytics looks for the root cause of a problem. It is use to determine why something happened. This type attempts to find and understand the causes of events and behaviors.

III. Predictive Analytics

Predictive analysis establishes previous data patterns and gives list of solutions, which may come for given situation. Predictive analyses study the present as well as past data and predict what may happen in future and give probabilities of what would happen. It is used to your big data to forecast other data, which we do not have. This analytical method is one of the most commonly methods used for sales lead scoring, social media and consumer relationship management data.

It consists of asking the question: What is likely to happen?

It uses past data in order to predict the future. It is all about forecasting. Predictive analytics uses many techniques like data mining and artificial intelligence to analyze current data and make scenarios of what might happen.

IV. Prescriptive Analytics

Prescriptive analysis reveals actions and recommend of what step should take. It gives answer to the situation in a focused way. Prescriptive data analytics goes one-step forward of predictive as it provides multiple actions with likely outcomes for each decision. This method of analytics is not preferred much by organizations, but it can show impressive result if used correctly.

It consists of asking the question: What should done?

It is dedicat to finding the right action to be taken. Descriptive analytics provides a historical data, and predictive analytics helps forecast what might happen. Prescriptive analytics uses these parameters to find the best solution.

After collected data, there are different tools use to analyze big data. In the next chapter,we will talk about two popular ones that use to analyze big data.

2.10 Benefits of Big Data Processing

Ability to process 'Big Data' brings in multiple benefits, such as-

- Businesses can utilize outside intelligence while taking decisions
- Improved customer service
- Early identification of risk to the services, if any
- Better operational efficiency.

2.11 Why is Big Data Important?

The importance of big data does not revolve around how much data a company has but how a company utilizes the collected data. Every company uses data in its own way; the more efficiently a company uses its data, potential it has to grow. The company can take data from any source and analyses it to find answers, which will enable:

Cost Savings: Some tools of Big Data like weka can bring cost advantages to business when large amounts of data are to be stored and these tool help in identifying more efficient ways of doing business.

Time Reductions: The high speed of can easily identify new sources of data, which helps businesses analyzing data immediately, and make quick decisions.

Understand the market conditions: By analyzing big data, you can get a better understanding of current market conditions.

Control online reputation: Big data tools can do sentiment analysis. Therefore, you can get feedback about who is saying what about your company. If you want to monitor and improve the online presence of your business, then, big data tools can help in all this.

Chapter 3

Statistical Method For analyzing Big Data

3.1 BRIEF OVERVIEW OF ANALYTICAL TOOLS

All data analysis tools have in common is the countless debates about why their programming language of choice is better, more advanced, faster, holier etc. In today's data science community, it seems as if these discussions are boundless with advocates of SAS, SPSS, R, Python, Julia, etc. battling and challenging each other on every online medium on the best statistical programming language .This overview gives us the brief knowledge about the tools so that afterwards comparison of these tools can be done and find out which is best among them. In this research, we talk about two statistical tools available for data analytics are brief as below.

3.2 SPSS

SPSS stands for Statistical Package for the Social Sciences. In 1979 SPSS jeopardized the University of Chicago's status as a tax-exempt organization. SPSS was acquired by IBM in 2009 for US$1.2 billion. SPSS was made to be easier to use then other statistical software like S Plus, R or SAS. SPSS is a great tool for non-statisticians since it has a user-friendly interface and easy to use drop down menus. Like Excel, SPSS is known beyond just the data science community. SPSS is primarily a statistical package, and offers a range of statistical tests, regression frameworks, correlations, and factor analyses. SPSS is a versatile package that allows many different types of analyses, data transformations, and forms of output - in short; it will more than adequately serve our purposes. SPSS is by far the easiest to learn. So if you only open a statistical program twice a month SPSS is the way to go.

3.3 WEKA

The WEKA workbench is a collection of machine learning algorithms and data preprocessing tools that include virtually all the algorithms are used. It is designed to quickly you can try out existing methods on new datasets in flexible ways. It provides extensive support for the whole process of experimental data mining, including preparing the input data, evaluating learning schemes statistically, and visualizing the input data and the result of learning.

As well as a wide variety of learning algorithms, it includes a wide range of preprocessing tools. This diverse and comprehensive toolkit is access through a common interface so that its users can compare different methods and identify those, which are most appropriate for the problem. WEKA was develop at the University of Waikato in New Zealand; the name stands for ***Waikato Environment for Knowledge Analysis***. Outside the university the WEKA, pronounced to rhyme with *Mecca*, is a flightless bird with an inquisitive nature found only on the islands of New Zealand. The system is written in Java and distributed under the terms of the GNU General Public License. It runs on almost any platform and has been tested under Linux, Windows, and Macintosh operating systems.

3.5 **What is in WEKA?**

WEKA provides implementations of learning algorithms that you can easily apply to your dataset. It also includes a variety of tools for transforming datasets, such as the algorithms for discretization and sampling. You can preprocess a dataset; feed it into a learning scheme and analyze the resulting classifier and its performance—all without writing any program code at all.

The workbench includes methods for the main data mining problems: regression, classification, clustering, association rule mining, and attribute selection. Getting to know the data is an integral part of the work, and many data visualization facilities and data preprocessing tools are provided. All algorithms take their input in the form of a single relational table that can be read from a file or generated by a database query.

One way of using WEKA is to apply a learning method to a dataset and analyze its output to learn more about the data. Another is to use learned models to generate predictions on new instances. A third is to apply several different learners and compare their performance in order to choose one for prediction.

In the interactive WEKA interface, you select the learning method you want from a menu. Many methods have tunable parameters, which you access through property sheet or *object editor*. A common evaluation module is used to measure the performance of all classifiers.

Implementations of actual learning schemes are the most valuable resource that WEKA provides. But tools for preprocessing the data, called *filters*, come a close second. Like classifiers, you select filters from a menu and tailor them to your requirements.

3.6 How do you use it?

The easiest way to use WEKA is through a graphical user interface called the *Explorer*. This gives access to all of its facilities using menu selection and form filling. For example, you can quickly read in a dataset from a file and build a decision tree from it. The Explorer guides you by presenting options as forms to be filled out. Helpful *tool tips* pop up as the mouse passes over items on the screen to explain what they do. Sensible default values ensure that you canget results with a minimum of effort but you will have to think about what you are doing to understand what the results mean.

There are three other graphical user interfaces to WEKA. The *Knowledge Flow* interface allows you to design configurations for streamed data processing. A fundamental disadvantage of the Explorer is that, it holds everything in main memory when you open a dataset, it immediately loads it all in. That means, it can only be applied to small-to medium-sized problems. However, WEKA contains some incremental algorithms that can be used to process svery large datasets. The Knowledge Flow interface lets you drag boxes representing learning algorithms and data sources around the screen and join them together into the configuration you want. It enables you to specify a data stream by connecting components representing data sources, preprocessing tools, learning algorithms, evaluation methods, and visualization modules. If the filters and learning algorithms are capable of incremental learning, data will be loaded and processed incrementally.

WEKA's third interface, the *Experimenter*, is designed to help you answer a basic practical question when applying classification and regression techniques: Which methods and parameter values work best for the given problem? There is usually no way to answer this question a priori, and one reason we developed the workbench was to provide an environment that enablesWEKA users to compare a variety of learning techniques. This can be done interactively using the Explorer. However, the Experimenter allows you to automate the process by making it easy to run classifiers and filters with different parameter settings on a corpus of datasets, collect performance statistics, and perform significance tests. Advanced users can employ the Experimenter to distribute the computing load across multiple machines using Java remote method invocation. In this way you can set up large-scale statistical experiments and leave them to run.

The fourth interface, called the *Workbench*, is a unified graphical user interface that combines the other three (and any plugging that the user has installed) into one application. The Workbench is highly configurable, allowing the user to specify which applications and plugins will appear, along with settings relating to them.

Behind these interactive interfaces lies the basic functionality of WEKA. This can be accessed in raw form by entering textual commands, which gives access to all features of the system. When you fire up WEKA you have to choose among five different user interfaces via the WEKAGUI Chooser: The Explorer, Knowledge Flow, Experimenter, Workbench, and command-line interfaces. Most people choose the Explorer, at least initially.

3.7 What else can you do?

An important resource when working with WEKA is the online documentation, which has automatically generate from the source code and concisely, reflects its structure. We will explain how to use this documentation. We will also identify WEKA's major building blocks, highlighting which parts contain supervised learning methods, which contain tools for data preprocessing, and which contain methods for other learning schemes. The online documentation gives the only complete list of available algorithms because WEKA is continually growing and being generat automatically from the source code the online documentation is always up to date. Moreover, it becomes essential if you want to proceed to the next level and access the library from your own Java programs or write and test learning schemes of your own.

In most data mining applications, the machine-learning component is just a small part of a far larger software system. If you intend to write a data mining application, you will wantto access the programs in WEKA from inside your own code. By doing so, you can solve the machine learning sub problem of your application with a minimum of additional programming.

3.8 Advantages of Weka include:

I. Free availability under the GNU General Public License.
II. Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
III. Comprehensive collection of data preprocessing and modeling techniques.
IV. Ease of use due to its graphical user interfaces.

3.9How do you get it?

WEKA is available from http://www.cs.waikato.ac.nz/ml/weka. You can download either a platform-specific installer or an executable Java jar file that you run in the usual way if Java to installed. We recommend that you download and install it.

3.10 Linear regression:

Modeling refers to the development of mathematical expressions that Describe in some sense the behavior of a random variable of interest. This

Variable may be the price of wheat in the world market, the number of deaths from lung cancer, the rate of growth of a particular type of tumor, or the tensile strength of metal wire. In all cases, this variable is called the **Dependent variable** and denoted with Y. A subscript on Y identifies the particular unit from which the observation was taken, the time at which the price was recorded, the county in which the deaths were recorded, the experimental unit on which the tumor growth was recorded, and so forth. Most commonly the modeling is aimed at describing how the **mean** of the dependent variable $E(Y)$ changes with changing conditions; the variance of the dependent variable is assumed to be unaffected by the changing Conditions. Other variables which are thought to provide information on the behavior of the dependent variable are incorporated into the model as predictor or explanatory variables. These variables are called the independent variables and are denoted by X with subscripts as needed to identify different independent variables. Additional subscripts denote the observational unit from which the data were taken. The X's are assumed to be known constants. In addition to the X's, all models involve unknown constants, called parameters, which control the behavior of the model.

The linear model is:

$$\boldsymbol{y_i = \beta_0 + \beta_1 X_i} + \varepsilon_i \boldsymbol{i = 1, 2, \ldots, n} \quad \textbf{.........(1)}$$

The random errors $\varepsilon_i$have zero mean and are assumed to have common Variance$\sigma^2$and to be pair wise independent. Since the only random element in the model is $\varepsilon_i$these assumptions imply that the *Yi* also have common variance $\sigma 2$ and are pairwise independent. For purposes of making tests of significance, the random errors are assumed to be normally distributed, which implies that the *Yi* are also normally distributed. The random error assumptions are frequently state as:

$$\varepsilon_i \sim NID(0, \sigma^2)$$

Where NID stands for "normally and independently distributed." The quantities in parentheses denote the mean and the variance, respectively, of the normal distribution

The simple linear model has two parameters β_0 andβ_1, which are to be Estimated from the data. If there were no random error in *Yi*, any two data Points can be used to solve explicitly for the values of the parameters. The random variation in *Y,* however, causes each pair of observed data Points to give different results.

The **least squares estimation procedure** uses the criterion that the solution must give the smallest possible sum of squared deviations of the Observed *Yi* from the estimates of their true means provided by the solution.

$$\boldsymbol{\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i i = 1, 2, \dots, n} \text{ ....(2)}$$

Let β_0 and β_1 be numerical estimates of the parameters $\widehat{\beta_0}$and$\widehat{\beta_1}$, calculated by:

$$\widehat{\boldsymbol{\beta_1}} = \frac{n\sum xy \quad \sum x \sum y}{n\sum x^2 - (\sum x)^2} \qquad \text{.........(3)}$$

and:

$$\widehat{\boldsymbol{\beta_0}} = \boldsymbol{\overline{y}} - \widehat{\boldsymbol{\beta_1}}\boldsymbol{\overline{x}} \text{ ....(4)}$$

The least squares principle chooses $\beta 0$ and _$\beta 1$ that minimize the sum of squares of the residuals.

Erroris the difference between an observed dependent value and one predicted from the regression equation:

$$\widehat{\varepsilon_i} = \boldsymbol{y_i} - \widehat{\boldsymbol{y_i}} = \boldsymbol{y_i} - (\widehat{\boldsymbol{\beta}}_0 + \widehat{\boldsymbol{\beta}}_1 \boldsymbol{X_i}) \qquad \textbf{.........(5)}$$

3.11 Error Measurement:

The following statistical indices are use to measure the model error.

1- Accuracy measures:

It includes estimates based on the calculation of the value *ei*

i. Mean absolute error (MAE):

The MAE used to measure the closeness of the prediction to the eventual outcomes, calculated by:

$$\mathrm{MAE} = \frac{1}{\mathrm{N}} \sum_{\mathrm{I}=1}^{\mathrm{N}} |\mathrm{y_i} - \widehat{\mathrm{y_i}}| \quad = \quad \frac{1}{\mathrm{N}} \sum_{\mathrm{I}=1}^{\mathrm{N}} |\mathrm{e_i}| \qquad \textbf{.........(6)}$$

ii. Root mean square error (RMSE)

While RMSE represents the sample standard deviation of the differences between predicted values $\hat{y}$and observed values,$\hat{y}$ where n is the number of observations and calculated by:

$$RMAE = \sqrt{\frac{1}{N}\sum_{I=1}^{N}(y_i - \hat{y}_i)^2} \qquad \textbf{.........(7)}$$

iii. Correlation

The correlation measures strength and direction of a linear relationship between and y. The value of R is always between +1 and −1 and the model optimality can be known by how close.

3.12 **Relative Measures:**

It is including:

i- Relative Mean Absolute Error:

Calculate by:

$$RRSE = \sqrt{\frac{\sum_{I=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{I=1}^{N}(\bar{y} - y_i)^2}} \qquad \textbf{.........(8)}$$

3 Relative Absolute Error

$$RAE = \frac{\sum_{I=1}^{N} |\hat{y}_i - y_i|}{\sum_{I=1}^{N} |\bar{y} - y_i|} \qquad \textbf{.........(9)}$$

The following represent the multiple linear regression model:

$$y_i + B_0 + B_1 X_i$$

Where Y is dependent variable, and $x_1, x_2, x_3, \ldots x_n,$ are independent variables. Also β0,β1,…,n are regression parameters, and $e_i$is an error of the model. To estimate the regression parameter vector B=(β0,β1,β2,…,βn) for the population, we extract a sample data set from the population, and compute an estimate parameter vector$\hat{B}$=($\hat{\beta}$0,$\hat{\beta}$1,$\hat{\beta}$2,…,$\hat{\beta}$n) using the sample. Hence, we can estimate the regression function as follow.

$$\hat{y}_i = \hat{B}_0 + \hat{B}_1 X_1 + \hat{B}_2 X_2 + \hat{B}_3 X_3 + \cdots + \hat{B}_n X_n$$

This is very standard approach in statistical analysis.in the big data analysis, we have new problem, which is different to the traditional statistics. This is to use and analyze whole data according to circumstances. In this research, we consider big data to the population of statistics, and separate the population into 20 subset.

Regression problems:

1- Autocorrelation:

The error terms are said to be auto correlated if and only if $cov(u_{i,} u_j)$ for I ≠ J

The error term at one date can correlatewith the error terms in the previous periods:

Autoregressive process of order k = 1, 2. . .

$$AR(k): u_t = p_1 u_{t1} + p_2 u_{t2} + p_3 u_{t3} + \cdots + p_k u_{tk} + v_t$$

Moving average process of order k $=$ $1, 2$, ...,

$$MA(k): u_t = v_t + \lambda_1 v_{t1} + \lambda_2 v_{t2} + \lambda_3 v_{t3} + \cdots + \lambda_k v_{tk}$$

(Cross $-$ section Data) The error terms may be correlated with each other in terms of socio and geographical distance such as the distancebetw een towns and neighborhood effects.

Assuming all other assumptions remain to hold, under the conditionof autocorrelation,the OLS estimator is still unbiased.

The OLS is not BLUE any more. The usual OLS standard errors and test statistics are no longer valid.
We can …nd an autocorrelation-robust estimator of the variance afterwe perform the OLS regression.Alternatively, we can devise an efficient estimator by reweighting thedata appropriately to take into account of autocorrelation.

We used Durbin-Watson test: based on the OLS residuals,

$$d = \frac{\sum_{t=2}^{N}(\widehat{u_t} - \widehat{u_{t-1}})^2}{\sum_{t=2}^{N}\widehat{u_t}^2} = 2(1 - r) + \frac{\widehat{u_1}^2 + \widehat{u_N}^2}{\sum_{t=2}^{N}\widehat{u_t}^2}$$

$$\text{Where } r = \frac{\sum_{t=2}^{N}(\widehat{u_t} \times \widehat{u_{t-1}})^2}{\sum_{t=2}^{N}\widehat{u_t}^2}$$

2- Multicollinearity:

Multicollinearity can be detect via various methods. In this article, we will focus on the most common one – **VIF (Variable Inflation Factors)**.

VIF determines the strength of the correlation between the independent variables. It is predict by taking a variable and regressing it against every other variable, or VIF score of an independent variable represents how well the variable is explained by other independent variables.

R^2Value is determined to find out how well an independent variable is describe by the other independent variables. A high value of R^2means that the variable is highly correlated with the other variables. This is capture by the **VIF** is denoted below:

$$\text{VIF} = \frac{1}{1-R^2}$$

Regression assumptions

Linear regression makes several assumptions about the data, such as :

1. **Linearity of the data**. The relationship between the predictor (x) and the outcome (y) assume linear.
2. **Normality of residuals**. The residual errors are assume normally distribute.
3. **Homogeneity of residuals variance**. The residuals are assumed to have a constant variance (**homoscedasticity**)
4. **Independence of residuals error terms**.

You should check whether these assumptions hold true. Potential problems include:

1. **Non-linearity** of the outcome - predictor relationships
2. **Heteroscedasticity**: Non-constant variance of error terms.
3. **Presence of influential values** in the data that can be:
   - Outliers: extreme values in the outcome (y) variable
   - High-leverage points: extreme values in the predictors (x) variable

All these assumptions and potential problems can check by producing some diagnostic plots visualizing the residual errors.

4.1 preface:

This chapter include the applied to what explained in the theoretical chapter and we will describe the data correlation **Root mean squared error RMSE** and **MAE** for simple linear regression and multiple liner regression **Root relative squared error** ,lastly **RMSE** and **Relative absolute error** to Comparative between estimation method ,

4.2 Describe of study's data:

The data linked to the volume of comments about a publication and the study data revolve around expected comments in a particular publication in order to predict the size of expected comments. The data relates to the number of (199030) comments.

Future values predicted based on existing values by regression analysis, as it is a statistical modeling method to find the relationship between the target variable (Y) and the prediction variables (X).[26]

B1 is the X coefficient and determines the rate of change in the target variable (Y) in one unit of variance in the forecast variable (X). The variables used to explain the target variable called interpretive variables or independent variables called a dependent variable. It mostly used to predict, predict in our case, explanatory variables are the features of Facebook pages, and the response variable is the size of the notes.

For this work, the concepts related to our domain are: (1) Source (independent variablex_2): source refers to the page that produces the post.

[26]Buza Krisztian, "Feedback Prediction for Blogs", Springer International Publishing on Data Analysis, Machine Learning and Knowledge

(2) Links (independent variablex_1): these are the pointers to other related posts or pages referred in main text or comments.

(3) Main text (independent variablex_3): the text refers to the main topic of the post. (4) Comments (dependent variabley): these are the opinions of the users about a post or other comments mentioned under the main text. (5) Feedback Volume (independent variablex_4): volume of feedback can be measure as the count of words in the comment section, the number of comments, the number of distinct users who leave comments, or a variety of other ways. These measures can be affected by various factors like main text of post, link to other posts, the time of day the post appears, a side conversation, Page likes, page check-ins, page talking about or page category etc.

We applied this study on generated data; it has independent normally distributed with different mean and different variance such as:

Table 1 Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Y	.463	199030	.030	.484	199030	.040

a. Lilliefors Significance Correction

The above table show test of normality for variable y the result of sig value when compare it with 0.05 show the variable y is belonging to normal distribution.

4.5Descriptive statistics of the simple regression data

We calculate means and stander deviation depending on the completed value of variables, to know is there ostensibly differences.

Table 2statistics results

	Minimum	maximum	Mean	Stander deviation
y	0	2119	21.815	74.658
x	1	106	24.24	19.9

From table (2)the results revealed that the mean of dependent variable Y is 21.8, mean of independent variables X is 24.2, and the stander deviation is 74.6 and 19.9 respectively.

4.6Mean absolute error results using weka:

To test the first hypotheses;is a statistically significant difference between simple regression equation in big data using weka program and spss package

We check these hypotheses by MAE and MSE

Table 2Mean absolute error results using weka

Data File No.	Sample no.	No of observation	Mean absolute error MAE
book001	1	1990	30.4416
book002	2	9951	31.0489
book003	3	29854	31.169
book004	4	39806	31.0761
book005	5	59709	30.8048
book006	6	69660	29.7427
book007	7	79612	28.6603
book008	8	89563	28.6198
book009	9	99515	26.5787
book010	10	109466	23.5662
book011	11	119418	21.5476
book012	12	129369	20.547
book013	13	139321	18.5613
book014	14	149272	18.6675
book015	15	159224	17.7821
book016	16	169175	15.9701
book017	17	179127	15.979

book018	18	189078	13.4933
book019	19	193059	12.441
book020	20	199030	10.542

Source: the researcher from applied study,weka package

From Table (3), the results revealed that when taking a different number of samples through the use of the measure of **Mean absolute error** and note that the value of the **Mean absolute error** was valued at 30.4416 at the size of the first sample 1990 and the researcher increased the size of the sample each time and with observation The value of the **Mean absolute error** noted decreased its value when increasing the size of the sample.

4.7Mean Square error results using weka:

Table 3Mean Square error results using weka

Data File No.	sample No.	No of observation (n)	mean squared error MSE
book001	1	1990	80.9793
book002	2	9951	81.4442
book003	3	29854	78.1668
book004	4	39806	78.0962
book005	5	59709	76.2927
book006	6	69660	74.4391
book007	7	79612	73.9182
book008	8	89563	71.8363
book009	9	99515	70.3982
book010	10	109466	70.206
book011	11	119418	68.1089
book012	12	129369	65.5772
book013	13	139321	64.5095

book014	14	149272	62.8355
book015	15	159224	60.6275
book016	16	169175	57.7028
book017	17	179127	56.9291
book018	18	189078	52.9879
book019	19	193059	49.1122
book020	20	199030	45.2544

Source: the researcher from applied study,weka package

From Table (3), the results revealed that when taking a different number of samples through the use of the measure of**mean square error**and note that the value of the **mean square** was valued at 80.9793 at the size of the first sample 1990 and the researcher increased the size of the sample each time and with observation The value of the **mean square**noted decreased its value when increasing the size of the sample.

The researcher test first hypothesis the researcher used spss package to the same data give the blow result.

4.8Mean absolute error results using spss:

Table5Mean absolute error results using spss

Data File No.	sample No.	No of observation (n)	Mean absolute errorMAE
book001	1	1990	30.44155
book002	2	9951	31.04887
book003	3	29854	30.189
book004	4	39806	30.8048
book005	5	59709	30.0761
book006	6	69660	29.7422
book007	7	79612	29.6603
book008	8	89563	28.6194
book009	9	99515	28.5788
book010	10	109466	28.2665
book011	11	119418	27.8476
book012	12	129369	27.7547
book013	13	139321	26.5613
book014	14	149272	26.6669
book015	15	159224	25.7821
book016	16	169175	24.9701
book017	17	179127	23.978
book018	18	189078	22.4910

book019	19	193059	21.440
book020	20	199030	20.544

Source: the researcher from applied study,spss package

From Table (4), the results revealed that when taking a different number of samples through the use of the measure of **Mean absolute error**and note that the value of the **Mean absolute error** was valued at 30.44155at the size of the first sample 1990 when we used spss, and the researcher increased the size of the sample each time and with observation The value of the **Mean absolute error** noted decreased its value when increasing the size of the sample, but it decreased slowly compression with weka.

4.9Mean Square error results using spss:

Table 4Mean Square error results using spss

Data File No.	**sample No.**	**No of observation (n)**	**mean squared error MSE**
book001	1	1990	80.98
book002	2	9951	81.44
book003	3	29854	80.022
book004	4	39806	80.117
book005	5	59709	80.045
book006	6	69660	79.112
book007	7	79612	79.100
book008	8	89563	78.877
book009	9	99515	78.401
book010	10	109466	78.067
book011	11	119418	77.100
book012	12	129369	76.814
book013	13	139321	76.068

book014	14	149272	75.335
book015	15	159224	75.005
book016	16	169175	74.551
book017	17	179127	73.851
book018	18	189078	72.118
book019	19	193059	71.794
book020	20	199030	70.987

Source: the researcher from applied study,spss package

From aboveTable the results revealed that when taking a different number of samples through the use of the measure of**mean squared error**and note that the value of the **mean squared** was valued at 80.9793 at the size of the first sample 1990 and the researcher increased the size of the sample each time and with observation The value of the **mean squared** noted decreased its value when increasing the size of the sample.

4.10Descriptive statistics of the multiple regression data

We calculate means and stander deviation depending on the completed value of variables, to know is there ostensibly differences.

Table 7statistics results

	Minimum	**maximum**	**Mean**	**Stander deviation**
y	0	2119	21.815	74
x_1	1	106	24.24	19.9
x_2	0	2495	485	538
x_3	0	2119	381	439
x_4	0	2095	380	430

From table (7)the results revealed that the mean of dependent variable Y is 21.8 and mean of independent variables Xis 24.2 , 485

,381 ,and 380 respectively and the stander deviation is 74.6 and 19.9 , 538 ,439, and 430 respectively.

To test the second hypotheses;is a statistically significant difference between multiple regression equation in big data using weka program and spss package

4.11Mean absolute error results for multiple regressions using weka:

Table 5Mean absolute error results for multiple regressions using weka

Data File No.	sample No.	No of observation (n)	Mean absolute errorMAE
book001	1	1990	26.5574
book002	2	9951	25.7376
book003	3	29854	23.6036
book004	4	39806	23.5748
book005	5	59709	21.3797
book006	6	69660	21.3682
book007	7	79612	20.3342
book008	8	89563	19.2709
book009	9	99515	17.2415
book010	10	109466	16.2835
book011	11	119418	15.2703
book012	12	129369	13.2902
book013	13	139321	12.3323
book014	14	149272	10.3435
book015	15	159224	9.4894
book016	16	169175	7.652
book017	17	179127	7.6828
book018	18	189078	5.2952

book019	19	193059	4.1552
book020	20	199030	3.2726

Source: the researcher from applied study,weka package

4.12Mean square error results for multiple regressions using weka:

Table 9Mean square error results

Data File No.	sample No.	No of observation (n)	Mean square errorMSE
book001	1	1990	77.0161
book002	2	9951	77.4792
book003	3	29854	74.0247
book004	4	39806	73.9735
book005	5	59709	72.3542
book006	6	69660	71.5281
book007	7	79612	71.0219
book008	8	89563	70.9428
book009	9	99515	68.5474
book010	10	109466	67.412
book011	11	119418	65.3299
book012	12	129369	63.8218
book013	13	139321	61.7412
book014	14	149272	58.071
book015	15	159224	56.8854
book016	16	169175	53.9758
book017	17	179127	50.1867
book018	18	189078	49.3224

book019	19	193059	45.4954
book020	20	199030	40.4676

Source: the researcher from applied study,weka package

4.13Mean absolute error results for multiple regressions using spss:

Table 6Mean absolute error results

Data File No.	sample No.	No of observation (n)	Mean absolute errorMAE
book001	1	1990	26.5574
book002	2	9951	25.7376
book003	3	29854	26.5574
book004	4	39806	25.650
book005	5	59709	24.6233
book006	6	69660	24.5232
book007	7	79612	23.511
book008	8	89563	23.9234
book009	9	99515	22.9299
book010	10	109466	22.8231
book011	11	119418	21.9266
book012	12	129369	21.7298
book013	13	139321	21. 235
book014	14	149272	20.9294
book015	15	159224	20.8226
book016	16	169175	20.6230
book017	17	179127	20.235
book018	18	189078	19.8230

book019	19	193059	19.6277
book020	20	199030	18.7222

Source: the researcher from applied study,spss package

From Table (4.10), the results revealed that when taking a different number of samples through the use of the measure of**mean absolute error**and note that the value of the **mean absolute** was valued at 26.6 at the size of the first sample 1990 and the researcher increased the size of the sample each time and with observation The value of the **mean absolute**noted decreased slowly its value when increasing the size of the sample.

4.14Mean square error results for multiple regressions using SPSS:

Table 11MSE results

Data File No.	**sample No.**	**No of observation (n)**	**mean squared error MSA**
book001	1	1990	77.0161
book002	2	9951	77.4792
book003	3	29854	77.0161
book004	4	39806	76.541
book005	5	59709	76.530
book006	6	69660	76.430
book007	7	79612	66.400
book008	8	89563	65.203
book009	9	99515	64.611
book010	10	109466	64.542
book011	11	119418	63.510
book012	12	129369	63.612
book013	13	139321	62.501

book014	14	149272	61.512
book015	15	159224	60.500
book016	16	169175	58.512
book017	17	179127	57.614
book018	18	189078	56.632
book019	19	193059	55.854
book020	20	199030	53.612

Source: the researcher from applied study,spss package

From Table (4.11), the results revealed that when taking a different number of samples through the use of the measure of**mean square error**and note that the value of the **mean square** was valued at 77.0161at the size of the first sample 1990 and the researcher increased the size of the sample each time and with observation The value of the **mean square**noted decreased its value when increasing the size of the sample. We used mean absolute error to determine accuracy of result, when we compare between result obtain by spss and weka result we note the mean absolute error decreasing slowly when using spss that is mean the accuracy of result obtain by weka butter than spss.

Mean Square error is necessary to remove any negative sings; it is gives more weight to large differences, the butter weka program than spss because give lower the value of MSE close to zero.

4.15Correlation coefficientresults:

To test thethird hypotheses;the correlation between big data variable is significant.

Table 12Correlation coefficient results

Data File No.	Sample No.	No of observation (n)	Correlation coefficient(R)
book001	1	1990	0.068
book002	2	9951	0.0798
book003	3	29854	0.0017
book004	4	39806	0.033
book005	5	59709	0.023
book006	6	69660	0.029
book007	7	79612	0.032
book008	8	89563	0.0350
book009	9	99515	0.0430
book010	10	109466	0.0489
book011	11	119418	0.0570
book012	12	129369	0.0600
book013	13	139321	0.0780
book014	14	149272	0.0799
book015	15	159224	0.0840
book016	16	169175	0.018
book017	17	179127	0.019

book018	18	189078	0.018
book019	19	193059	0.0166
book020	20	199030	0.019

Source: the researcher from applied study,weka package

From Table (4.12), the results show that when taking a different number of samples through the use of the measure is**Correlation coefficient**and note that the value of the **correlation coefficient** was valued at 0.0686 at the size of the first sample 1990 and the researcher increased the size of the sample each time and with observation The value of the **correlation coefficient**noted increased its value when increasing the size of the sample.

Table 13Correlation coefficient results using spss

Data File No.	**sample No.**	**No of observation (n)**	**Correlation coefficient (R)**
book001	1	1990	0.0685
book002	2	9951	0.0797
book003	3	29854	0.0817
book004	4	39806	0.0833
book005	5	59709	0.0910
book006	6	69660	0.120
book007	7	79612	0.220
book008	8	89563	0.250
book009	9	99515	0.270
book010	10	109466	0.289
book011	11	119418	0.320
book012	12	129369	0.360
book013	13	139321	0.390
book014	14	149272	0.429

book015	15	159224	0.480
book016	16	169175	0.512
book017	17	179127	0.560
book018	18	189078	0.577
book019	19	193059	0.583
book020	20	199030	0.616

Source: the researcher from applied study,spss package

From Table (4.13), the results revealed that when taking a different number of samples through the use of the measure of**Correlation coefficient**and note that the value of the **correlation coefficient** was valued at 0.0686 at the size of the first sample 1990 and the researcher increased the size of the sample each time and with observation The value of the **correlation ccoefficient**noted increased its value when increasing the size of the sample.

4.15**Relative Measures results:**

Table 14Relative absolute error using weka

Data File No	Sample no.	Relative absolute errorRAE
book001	1	90.4131
book002	2	90.7395
book003	3	91.3285
book004	4	91.2405
book005	5	91.5554
book006	6	93.5361
book007	7	93.5011
book008	8	94.5258
book009	9	94.5319
book010	10	95.5363
book011	11	96.5328
book012	12	96.543
book013	13	97.5095
book014	14	97.5648
book015	15	98.5618
book016	16	98.5811
book017	17	98.5518
book018	18	99.5129
book019	19	99.4944
book020	20	99.5495

Source: the researcher from applied study,weka package

From Table (4.18), the results revealed that when taking a different number of samples through the use of the measure of**Relative absolute**

errorand note that the value of the **Relative absolute error**was valued at 90%at the size of the first sample 1990 and the researcher increased the size of the sample each time and with observation The value of the **Relative absolute error**noted increased its value when increasing the size of the sample.

Table 15Root relative square error

Data File No.	Sample no.	Root relative square error RRSE
book001	1	90.7992
book002	2	90.6888
book003	3	91.4819
book004	4	91.467
book005	5	91.4806
book006	6	92.4712
book007	7	92.4515
book008	8	93.4565
book009	9	93.4569
book010	10	95.456
book011	11	96.4553
book012	12	96.447
book013	13	97.4443
book014	14	97.4562
book015	15	98.4564
book016	16	98.4607
book017	17	98.4594
book018	18	99.4661
book019	19	99.4701
book020	20	99.4592

Source: the researcher from applied study,weka package

From Table (4.15), the results revealed that when taking a different number of samples through the use of the measure of**Root relative squared error**and note that the value of the **Root relative squared**

errorwas valued at 90%at the size of the first sample 1990 and the researcher increased the size of the sample each time and with observation The value of the **Root relative squared error**noted increased its value when increasing the size of the sample.

Table 16Relative absolute error

Data File No.	Sample No.	Relative absolute error RAE
book001	1	83.7282
book002	2	83.8902
book003	3	84.7797
book004	4	84.8658
book005	5	85.0764
book006	6	85.2589
book007	7	85.2976
book008	8	85.1844
book009	9	85.2248
book010	10	86.3937
book011	11	86.4199
book012	12	86.3807
book013	13	86.4829
book014	14	87.2766
book015	15	88.4457
book016	16	89.541
book017	17	91.565
book018	18	91.6333
book019	19	92.8835
book020	20	93.4356

Source: the researcher from applied study,spss package

From Table (4.16), the results revealed that when taking a different number of samples through the use of the measure of**Relative absolute**

error RAEand note that the value of the **Relative absolute error RAE** was valued at 90%at the size of the first sample 1990 and the researcher increased the size of the sample each time and with observation The value of the **Relative absolute error RAE** noted increased slowly its value when increasing the size of the sample.

Table 7Root relative squared error using spss

Data File No.	Sample no.	Root relative squared error RRSE
book001	1	95.1059
book002	2	94.8356
book003	3	94.2103
book004	4	94.2162
book005	5	94.3425
book006	6	94.3131
book007	7	94.2798
book008	8	94.2822
book009	9	94.3091
book010	10	94.3709
book011	11	94.3838
book012	12	94.3685
book013	13	95.3422
book014	14	95.3846
book015	15	96.3994
book016	16	97.4316
book017	17	97.423
book018	18	98.5453
book019	19	98.6332
book020	20	99.387

Source: the researcher from applied study,SPSS package

From Table (17), the results revealed that when taking a different number of samples through the use of the measure of**Root relative square error** and note that the value of the **Root relative square error**was valued at 90%at the size of the first sample 1990 and the researcher increased the size of the sample each time and with observation The value of the **Root relative**

square error noted increased its value when increasing slowly the size of the sample.

Autocorrelation problem results using spss:

The researcher check Autocorrelation problem by using spss program and obtain the bellow results

Table 8Durbin-Watson test using spss

Data File No.	**Sample no.**	**No of observation**	D_U	D_L
book001	1	1990	2.006	0.091
book002	2	9951	2.008	0.091
book003	3	29854	2.067	0.096
book004	4	39806	2.090	0.094
book005	5	59709	2.086	0.097
book006	6	69660	2.031	0.093
book007	7	79612	3.006	0.095
book008	8	89563	3.003	0.095
book009	9	99515	3.018	0.094
book010	10	109466	3.023	0.094
book011	11	119418	3.031	0.094
book012	12	129369	3.034	0.093
book013	13	139321	3.036	0.092
book014	14	149272	3.039	1.09
book015	15	159224	3.041	1.092
book016	16	169175	4.041	1.095
book017	17	179127	4.04	2.093
book018	18	189078	4.040	2.090

book019	19	193059	4.041	2.091
book020	20	199030	4.040	2.095

Source: the researcher from applied studyspss, package

Autocorrelation problem results using weka:

The researcher check Autocorrelation problem by using weka program and obtain the bellow results

Table 9Durbin-Watson test using weka

Data File No.	**Sample no.**	**No of observation**	D_U	D_L
book001	1	1990	2.008	0.098
book002	2	9951	2.007	0.098
book003	3	29854	2.006	0.097
book004	4	39806	2.005	0.096
book005	5	59709	2.005	0.096
book006	6	69660	2.004	0.095
book007	7	79612	3.006	0.095
book008	8	89563	3.003	0.095
book009	9	99515	3.018	0.094
book010	10	109466	3.023	0.094
book011	11	119418	3.031	0.094
book012	12	129369	3.034	0.093
book013	13	139321	3.036	0.092
book014	14	149272	3.039	0.092
book015	15	159224	3.041	0.092
book016	16	169175	3.040	0.090
book017	17	179127	3.043	0.090

book018	18	189078	3.040	0.090
book019	19	193059	3.046	0.091
book020	20	199030	3.048	0.090

Source: the researcher from applied study,weka package

The Durbin-Watson test statistic tests the null hypothesis that the residuals from an ordinary least-squares regression are not autocorrelated against the alternative that the residuals follow an AR1 process. The Durbin-Watson statistic ranges in value from 0 to 4,from above data when we do compression between result obtain from weka and spss the weka give better result than spss.

Multicollinearityproblem results using spss:

The researcher check Autocorrelation problem by using spss program and obtain the bellow results

Table 10Multicollinearity test using spss

Data File No.	Sample no.	No of observation	VIF
book001	1	1990	0.03
book002	2	9951	0.09
book003	3	29854	0.1
book004	4	39806	1.2
book005	5	59709	3.5
book006	6	69660	4.03
book007	7	79612	4.99
book008	8	89563	5.6
book009	9	99515	5.98
book010	10	109466	7.002
book011	11	119418	8.01
book012	12	129369	9.35
book013	13	139321	9.89
book014	14	149272	10.01
book015	15	159224	10.52
book016	16	169175	11.9
book017	17	179127	13.6
book018	18	189078	13.9
book019	19	193059	14.5

book020	20	199030	15.077

Source: the researcher from applied study, SPSS package

From aboveTable , the results revealed that when taking a different number of samples through the use of the measure of Multicollinearity using VIF note that the value of the VIF was valued at between 0.03 and 15 at the size of the first sample 1990 and 199030.

Multicollinearity problem results using weka:

The researcher check Autocorrelation problem by using weka program and obtain the bellow results

Table 11Multicollinearity test using weka

Data File No.	Sample no.	No of observation	VIF
book001	1	1990	0.03
book002	2	9951	0.09
book003	3	29854	0.015
book004	4	39806	0.094
book005	5	59709	0.6
book006	6	69660	0.7
book007	7	79612	0.88
book008	8	89563	0.1
book009	9	99515	0.29
book010	10	109466	1.001
book011	11	119418	1.085
book012	12	129369	3.082
book013	13	139321	3.084
book014	14	149272	4.083
book015	15	159224	5.080
book016	16	169175	5.09
book017	17	179127	6.081

book018	18	189078	6.080
book019	19	193059	6.079
book020	20	199030	7.077

Source: the researcher from applied study,weka package

From Table , the results revealed that when taking a different number of samples through the use of the measure of Multicollinearity using VIF note that the value of the VIF was valued at between 0.03 and 7 at the size of the first sample 1990 and 199030.
From the above result weka butter than spss in solving Multicollinearity problem. Because VIF increases, the less reliable your regression results are going to be. In general, a VIF above 10 indicates high correlation and is cause for concern.

Regression diagnostics

Diagnostic plots Regression diagnostics plots can create using the weka base function

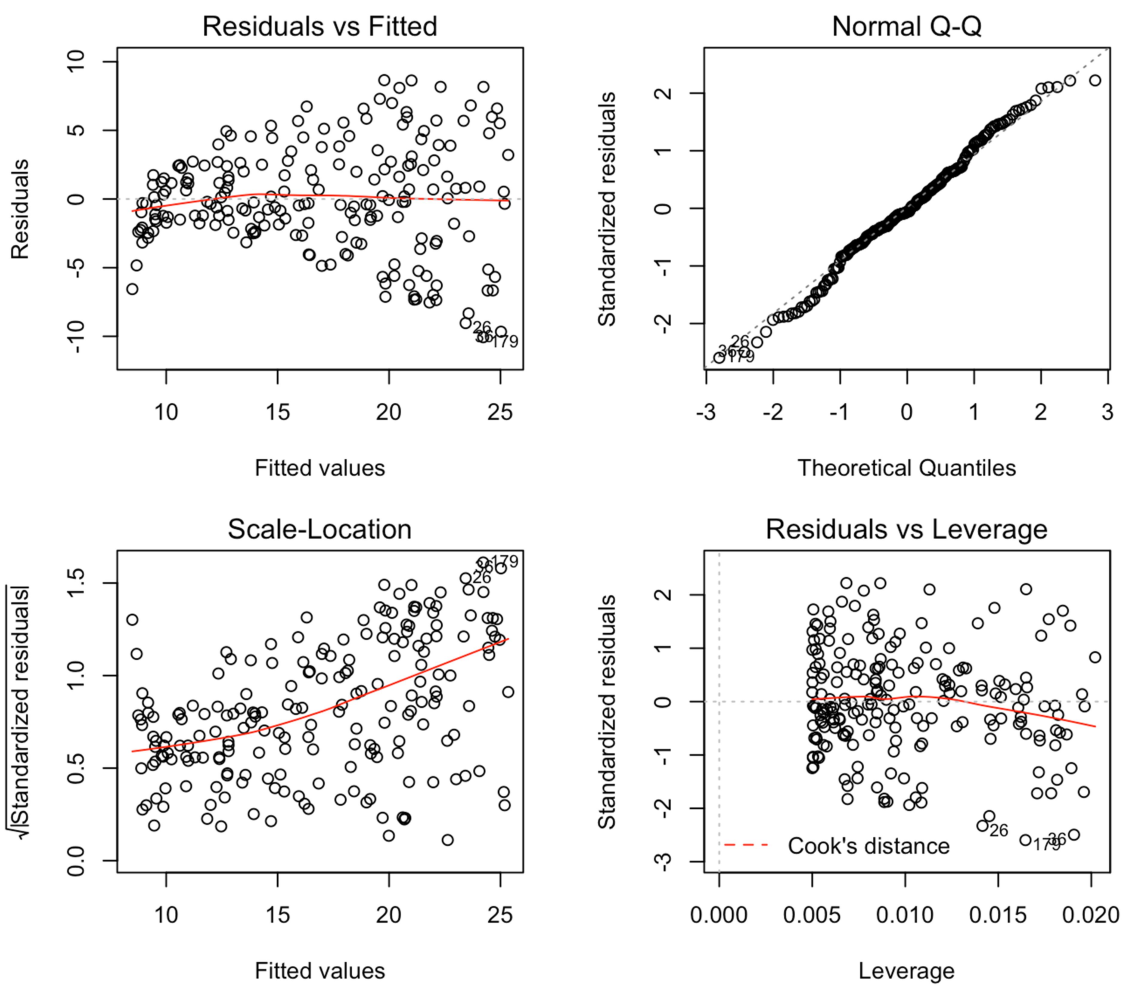

Figure 1Diagnostic plots Regression

The diagnostic plots show residuals in four different ways:

1. **Residuals vs Fitted**. Used to check the linear relationship assumptions. A horizontal line, without distinct patterns is an indication for a linear relationship, what is good.
2. **Normal Q-Q**. Used to examine whether the residuals are normally distributed. It is good if residuals points follow the straight dashed line.

3. **Scale-Location** (or Spread-Location). Used to check the homogeneity of variance of the residuals (homoscedasticity). Horizontal line with equally spread points is a good indication of homoscedasticity. This is not the case in our example, where we have a heteroscedasticity problem.
4. **Residuals vs Leverage**. Used to identify influential cases that extreme values might influence the regression results when included.

In the following section, we will describe, in details, how to use these graphs and metrics to check the regression assumptions and to diagnostic potential problems in the model.

Linearity of the data

The linearity assumption can check by inspecting the **Residuals vs fitted** plot

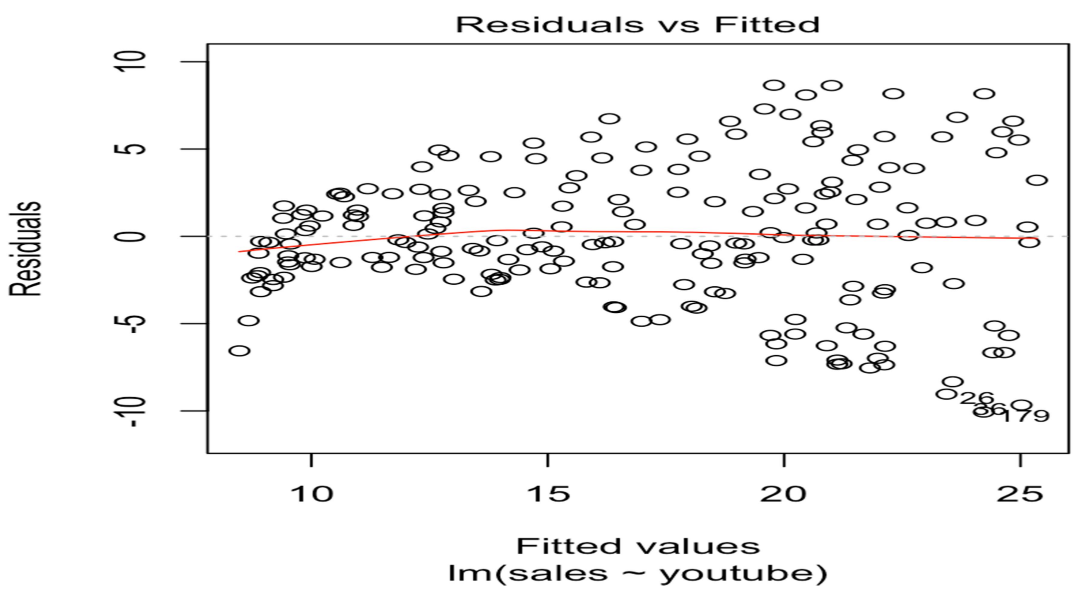

Figure 2Linearity of the data

Ideally, the residual plot will show no fitted pattern. That is, the red line should be approximately horizontal at zero. The presence of a pattern may indicate a problem with some aspect of the linear model.

There is no pattern in the residual plot. This suggests that we can assume linear relationship between the predictors and the outcome variables

Homogeneity of variance

This assumption can check by examining the *scale-location plot*, also known as the *spread-location plot*.

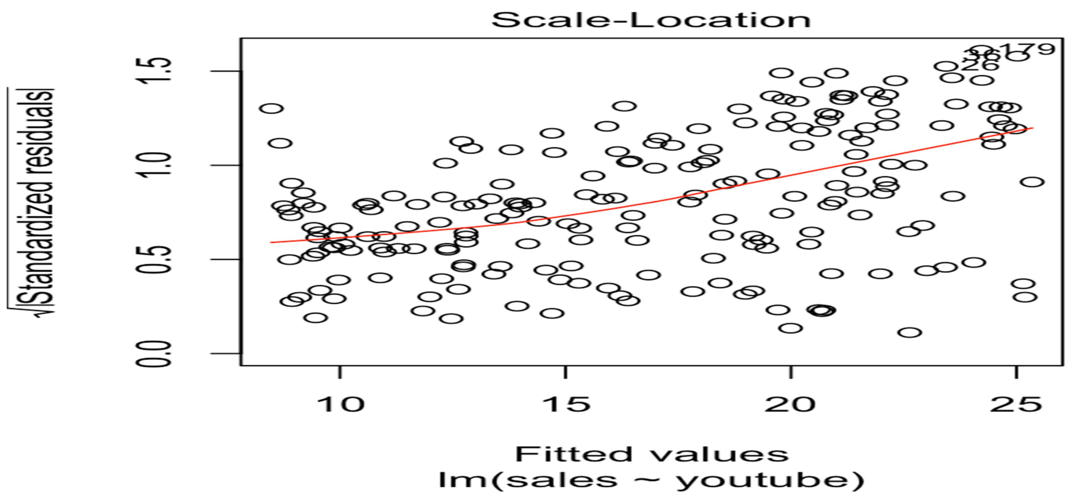

Figure 3Homogeneity of variance

This plot shows if residuals are spread equally along the ranges of predictors. It is good if you see a horizontal line with equally spread points. In our example, this is not the case.

It can be seen that the variancesof the residual points increases with the value of the fitted outcome variable, suggesting non-constant variances in the residuals errors (or *heteroscedasticity*).
A possible solution to reduce the heteroscedasticity problem is to use a log or square root transformation of the outcome variable (y).

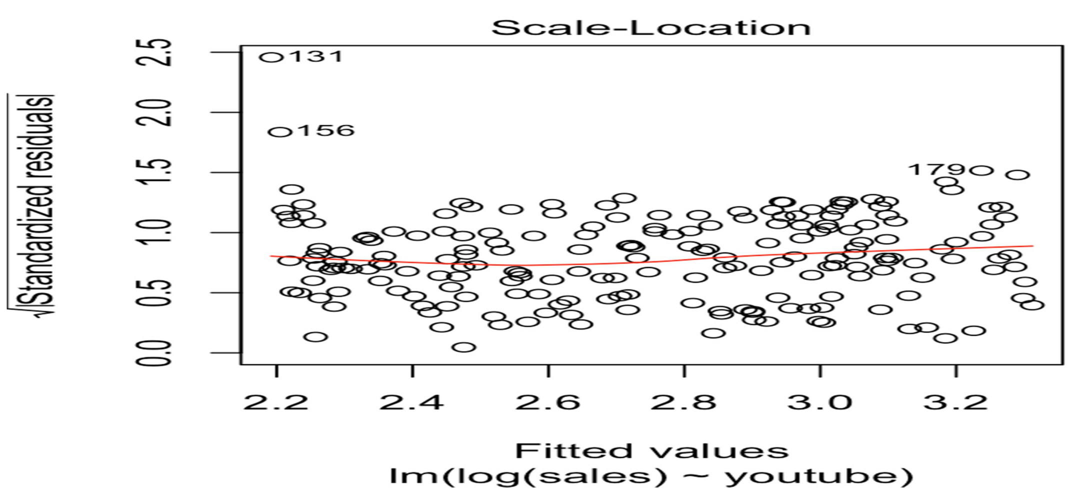

Figure 4Homogeneity of variance solutions

Normality of residuals

The QQ plot of residuals can use to visually check the normality assumption. The normal probability plot of residuals should approximately follow a straight line.

In our example, all the points fall approximately along this reference line, so we can assume normality.

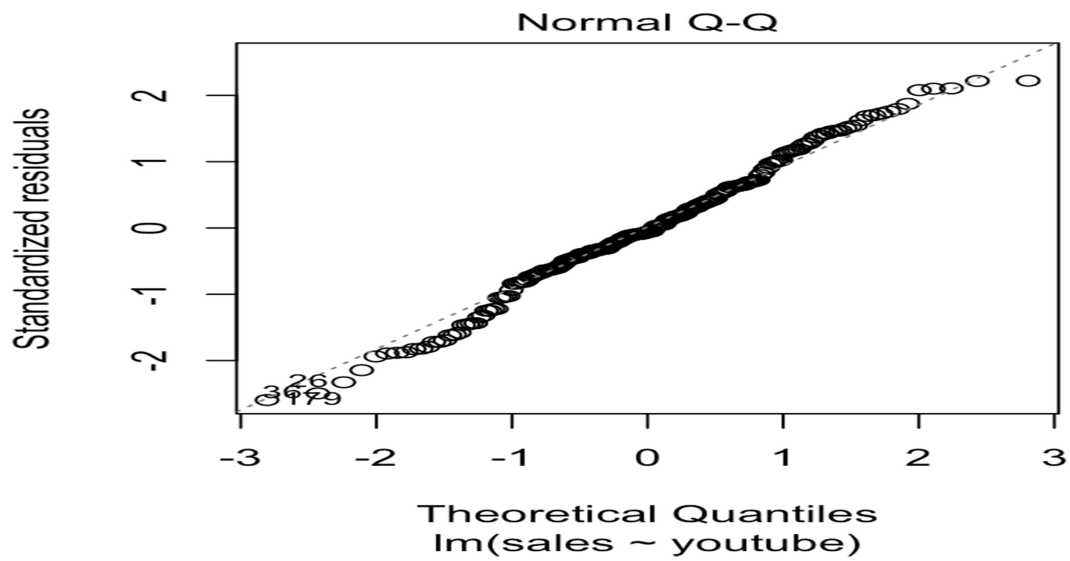

Figure 5Normality of residuals

Outliers and high leverage points

Outliers:
An outlier is a point that has an extreme outcome variable value. The presence of outliers may affect the interpretation of the model, because it increases the RSE.

Outliers can be identified by examining the *standardized residual* (or *studentized residual*), which is the residual divided by its estimated standard error. Standardized residuals can be interpreted as the number of standard errors away from the regression line.
Observations whose standardized residuals are greater than 3 in absolute value are possible outliers

High leverage points:
A data point has high leverage, if it has extreme predictor x values. This can be detected by examining the leverage statistic or the *hat-value*. A value of this statistic above $2(p + 1)/n$ indicates an observation with high leverage (P. Bruce and Bruce 2017); where, p is the number of predictors and n is the number of observations.

Outliers and high leverage points can be identified by inspecting the *Residuals vs Leverage* plot:

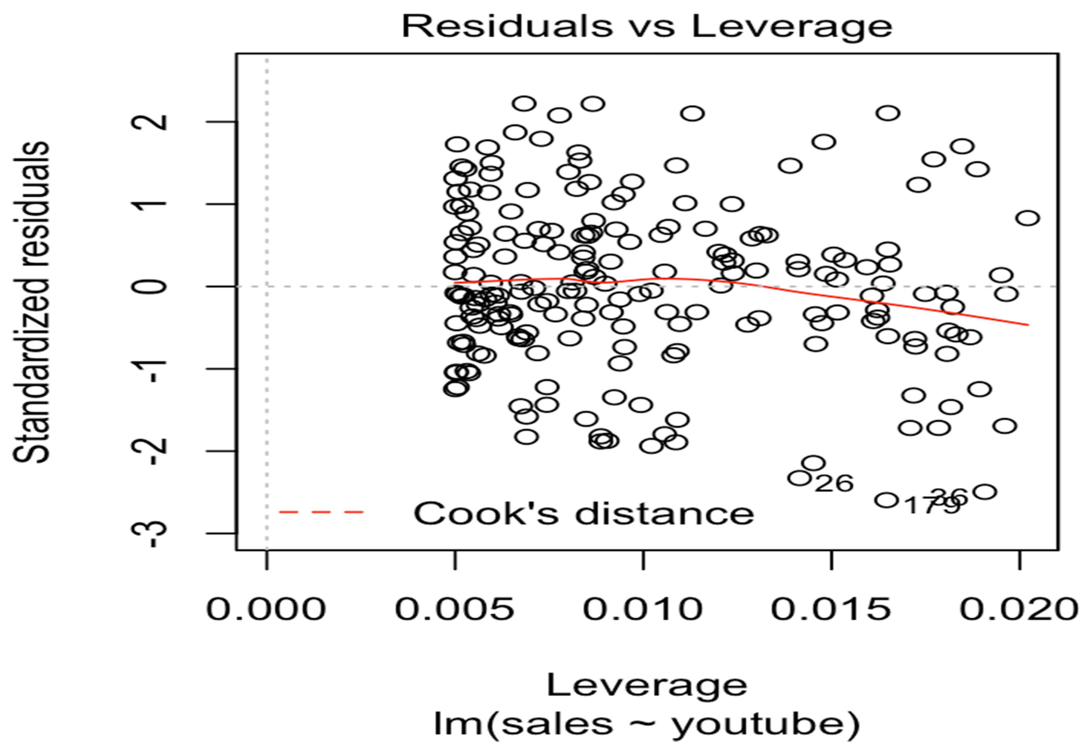

Figure 6Outliers and high leverage points

The plot above highlights most extreme points (#26, #36 and #179), with a standardized residuals below -2. However, there is no outliers that exceed 3 standard deviations, Additionally, there is no high leverage point in the data. That is, all data points, have a leverage statistic below 2(p + 1)/n = 4/200 = 0.02.

Chapter 4

Results and Recommendations

5.2 Results:

Result obtained by the first case; the regression equation obtained by WEKA program is more relative efficiency than SPSS package.

Results obtained by WEKA program, all value of Mean absolute error results decreasing quickly when increasing the data comparison with SPSS decreasing slowly.

In addition,all value of Mean square error results decreasing quickly when increasing the data comparison with SPSS program.

The results revealed that there is no ostensibly a difference between means and variance, Results obtained by correlation; WEKA gives butter correlation when increasing sample size

In WEKA program WEKA solve problem of regression autocorrelation and Multicollinearity and gives better result than SPSS.

5.3 **Recommendations:**

This study is recommend the following:

More attention should paid to the big data in the regression and performance of any study, theApplication of the above package in the different size of data is very important in statistics.

Using WEKA program in big data regression is better than other applications in producing equation that is more efficient.

5.3 REFERENCES

1. Buza Krisztian, 2014 "Feedback Prediction for Blogs", Springer International Publishing on Data Analysis, Machine Learning and Knowledge Discovery, pp. 145-152.
2. M.Tsagkias, W. Weerkamp, M. de Rijke, 2009 "Predicting the Volume of Comments on Online News Stories",CIKM"09 Proceedings of the 18th ACM conference on Information and knowledge management , pp.1765-1768.
3. Balali, A. and Rajabi, A. and Ghassemi, S. andAsadpour, M. and Faili, H., "Content diffusion prediction in social networks", Information and Knowledge Technology (IKT), 5th IEEE Conference, 2013, pp. 467-471. doi: 10.1109/IKT.2013.6620114
4. M. Tsagkias, W. Weerkamp, M. de Rijke, "News Comments: Exploring, Modeling, and Online Prediction" proceedings of the 32nd European conference.
5. Jamali, S. and Rangwala, H.2009, "Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis", Web Information Systems and Mining, IEEE International Conference, pp. 32-38. doi: 10.1109/WISM.2009.15.
6. Yano, Tae, and Noah A. Smith. "What's Worthy of Comment? Content and Comment Volume in Political Blogs" In 4th International AAAI Conference on Weblogs and Social Media, 2010.
7. Negi, S. and Chaudhury, S., 2012 "Predicting User-to-content Links in Flickr Groups", Advances in Social Networks Analysis and Mining (ASONAM), IEEE/ACM International Conference.
8. Rong Zhang, Zhenjie Zhang, Xiaofeng He, Aoying Zhou, "Dish Comment Summarization Based on Bilateral Topic Analysis" Data Engineering (ICDE), 31st IEEE International Conference, 2015, pp. 483 – 494. doi: 10.1109/ICDE.2015.7113308
9. H.Wimmer & L. M. Powell (2015). "A Comparison of Open Source Tools for Data Science". Proceedings of the Conference on Information Systems Applied Research Wilmington, North Carolina USA
10. Exploring, Modeling, and Online Prediction", ECIR'2010 Proceedings of the 32nd European conference on Advances in Information Retrieval, Springer

11. Ishwarappa, and J. Anuradha, “A Brief Introduction On Big Data 5Vs Characteristics And Hadoop Technology”. Procedia Computer Science 48, pp. 319-324, 2015.

12. Negi, S. and Chaudhury, S., “Predicting User-to-content Links in Flickr Groups”, Advances in Social Networks Analysis and Mining (ASONAM), IEEE/ACM International Conference, 2012 pp. 124-131. doi: 10.1109/ASONAM.2012.31.

13. Dr. K. J. Begum & Dr. A. Ahmed, “The Importance of Statistical Tools in Research Work”. International Journal of Scientific and Innovative Mathematical Research (IJSIMR)

14. Rahman, M.M. , “Intellectual knowledge extraction from online social data”, Informatics, Electronics Vision (ICIEV), IEEE International Conference, 2012, pp. 205-210. doi:10.1109/ICIEV.2012.6317392

15. Rong Zhang, Zhenjie Zhang, Xiaofeng He, Aoying Zhou, “Dish Comment Summarization Based on Bilateral Topic Analysis” Data Engineering (ICDE), 31st IEEE International Conference, 2015, pp. 483 – 494. doi: 10.1109/ICDE.2015.7113308

16. Buza Krisztian, “Feedback Prediction for Blogs”, Springer International Publishing on Data Analysis, Machine Learning and Knowledge Discovery,2014

17. Balali, A. and Rajabi, A. and Ghassemi, S. andAsadpour, M. and Faili, H., “Content diffusion prediction in social networks”, Information and Knowledge Technology (IKT), 5th IEEE Conference

18. Analytics Vidhya https://www.analyticsvidhya.com/blog/2014/03/sas-vs-vs-python-tool-learn/

19. Yano, Tae, and Noah A. Smith. “What's Worthy of Comment? Content and Comment Volume in Political Blogs”

20. K. Rangra ,Dr. K. L. Bansal, “Comparative Study of Data Mining Tools”, International Journal of Advanced Research in Computer Science and Software Engineering

21. A.Komathi, T.Ramya, M. Shanmugapriya, V. Sarmila, “A Novel Comparative Study on Data Mining Tools.

22. M.Tsagkias, W. Weerkamp, M. de Rijke, “Predicting the Volume of Comments on Online News Stories”,CIKM‟09 Proceedings of the 18th ACM conference on Information and knowledge management

23. M.Tsagkias, W. Weerkamp, M. de Rijke, “Predicting the Volume of Comments on Online News Stories”,CIKM‟09 Proceedings of the 18th ACM conference on Information and knowledge
24. Jamali, S. and Rangwala, H., “Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis
25. Dr. K. J. Begum & Dr. A. Ahmed, “The Importance of Statistical Tools in Research Work”. International Journal of Scientific and Innovative
26. H. Wimmer & L. M. Powell (2015). “A Comparison of Open Source Tools for Data Science“. Proceedings of the Conference on Information Systems Applied Research Wilmington, North Carolina USA. 2167-1508: v8 n3651.
27. Davenport, T. H., & Patil, D. (2012). Data scientist. Harvard Business Review, 90, 7076.
28. Dr. K. J. Begum & Dr. A. Ahmed, “The Importance of Statistical Tools in Research Work”. International Journal of Scientific and Innovative Mathematical Research (IJSIMR) Volume 3, Issue 12, December 2015, PP 50-58
29. https://www.analyticsvidhya.com/blog/2014/03/sas-vs-vs-python-tool-learn/
30. Office of Research Integrity, https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dhtopic.html
31. Data Camp, https://www.datacamp.com/community/tutorials/statistical language-wars-the-infograph#gs.9dggy7w
32. K. Rangra ,Dr. K. L. Bansal, “Comparative Study of Data Mining Tools”, International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 6, June 2014.
33. A.Komathi, T.Ramya, M. Shanmugapriya, V. Sarmila, “A Novel Comparative Study on Data Mining Tools”, International Journal of Innovative Research in Computer and Communication Engineering Vol. 4, Issue 11, November 2016
34. CRAN-R, https://cran.r-project.org/
35. TIOBE Software Index (2011). "TIOBE Programming Community Index Python".
36. "Programming Language Trends - O'Reilly Radar". Radar.oreilly.com. 2 August 2006.

37. "The RedMonk Programming Language Rankings: January 2011 – tecosystems". Redmonk.com.
38. Masoud Nosrati, “Python: An appropriate language for real world programming”, World Applied Programming, Vol (1), No (2), June 2011.
39. Goodger, David. "Code Like a Pythonista: Idiomatic Python"
40. Data Camp, https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis
41. http://www.fundinguniverse.com/company-histories/spss-inc-history
42. Arun menachery, “INTRODUCTION TO SPSS”. Data Camp,
43. https://www.datacamp.com/community/tutorials/statistical-language-wars-the-infograph#gs.N3o9EvM
44. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutmann, P., & Witten, I. (2009). The WEKA Data Mining Software: An Update. ACM SIGKDD Explorations Newsletter, 11(1), 10-18. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summa ry?doi=10.1.1.148.3671
45. Witten, I.H., Frank, E.: “Data Mining: Practical machine Learning tools and techniques”, 2nd addition,Morgan Kaufmann, San Francisco(2005).
46. https://stats.stackexchange.com/questions/33780/r-vs-sas-why-is-sas-preferred-by-private-companies
47. http://analyticstraining.com/2012/pricing-for-analytical-tools-in-india
48. https://www.21stcenturyleaders.org/why-is-community-service-important/
49. Marcus R. Wigan, Roger Clarke, “Big Data’s Big Unintended Consequences” Published by the IEEE Computer Society, pp 46-53
50. Srivastava S. (2017) Novel Method for Predicting Academic Performance of Students by Using Modified Particle Swarm Optimization (PSO). In: Panigrahi B., Hoda M., Sharma V., Goel S. (eds) Nature Inspired Computing. Advances in Intelligent Systems and Computing, vol 652. Springer, Singapor
51. joseph Hellerstein. The commoditization of massive data analysis. Blog on O’Reilly.com, 19 November 200
52. Big data, big impact: new possibilities for international development. World Economic Forum, 2012

53. James Manyika and others. Big data: the next frontier for innovation, competition and productivity. McKinsey Global Institute, May 2011.

54. Danah Boyd and Kate Crawford. Six provocations for Big Data. A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 2011

List of Tables

List of figures