A Regression Analysis on the Flow of EMU Library USERS

Ademola Ezekiel Babalola

Submitted to the
Institute of Graduate Studies and Research
in partial fulfilment of the requirements for the degree of

Master of Science
in
Applied Mathematics and Computer Science

Eastern Mediterranean University
February 2016
Gazimağusa, North Cyprus

Approval of the institute of Graduate Studies and Research

Prof. Dr. Cem Tanova
Acting Director

I certify that this thesis satisfies the requirement as a thesis for the degree of Master of Science in Applied Mathematics and Computer Science.

Prof. Dr. Nazim Mahmudov
Chair, Department of Mathematics
and Computer Science

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Master of Science in Applied Mathematics and Computer Science.

Asst. Prof. Dr Mehmet Ali Tut
Supervisor

Examining Committee

1. Prof. Dr. Rashad Aliyev
2. Asst. Prof. Dr Ersin Bodur
3. Asst. Prof. Dr Mehmet Ali Tut

ABSTRACT

This paper highlights the importance of statistics in everyday life in this 21st century. Statistics which is based on collecting, managing, processing and disseminating information, it applications are seen in Banks, Airports, information technology and schools.

This research presents a simple linear regression model, its derivation and how it is used to analyse data's obtained from the school library. The data's which defines the average population of students who use the library yearly. The model obtained is used to predict future use of the library by students, as well as an estimate of students that used the library before 2007. A statistical software application, called SPSS was used in the analysis.

The result obtained from this research shows about 33% of data can be analyze which makes our model somewhat a good fit for the data; however a non-linear model would be best to describe the library data.

Keywords: linear regression models, Sample data, correlation coefficient, influential point, SPSS.

ÖZ

Bu yazıda, bu 21. yüzyılda gündelik hayatın istatistiğin önemini vurgulamaktadır. Toplama, yönetme, işlenmesi ve bilginin yayılmasından dayanmaktadır İstatistik, bu uygulamalar Bankalar, Havaalanları, bilgi teknolojileri ve okullarda görülür.

Bu araştırma, basit bir doğrusal regresyon modeli, kendi türetme sunar ve okul kütüphanesinden elde edilen veriler 's analiz nasıl kullanılır. Yıllık kütüphane kullanımı öğrencilerin ortalama nüfus tanımlayan veri en. Elde edilen model, gelecek öğrenciler tarafından kütüphane kullanımı yanı sıra analizde kullanılan SPSS denilen 2007. istatistik programı uygulamadan önce kütüphaneyi kullanılan öğrenciler bir tahmin, tahmin etmek için kullanılır.

Bu araştırmada elde edilen sonuç verilerinin yaklaşık% 33 modelimizi verileri için biraz iyi bir uyum kılan analiz olabilir göstermektedir; Ancak doğrusal olmayan bir model kütüphanesi verilerini açıklamak için iyi olurdu.

Anahtar Kelimeler: Doğrusal regresyon modelleri, örnek veriler, korelasyon katsayısı, etkili nokta, SPSS.

DEDICATION

To the glory of God, i dedicate this thesis to my parent and siblings.

ACKNOWLEDGEMENT

I would to appreciate my Supervisor Asst. Prof. Dr Mehmet Ali TUT for his timely effort and tutoring through my research writing, for creating time out of his tight schedule just to ensure i meet up to target. He gave me a lot of insight encouragement and advices for the future.

I also cannot forget to mention Asst. Prof. Dr Yucel Tandogdu for his tutoring skills and advices off and on academics level. To every member of the committee I appreciate all timely effort and advices; I was thrilled and encouraged on how you skilfully guided me through my research study.

Many thanks to my Parent Revd. Canon. David Babalola and Mrs Olutoke Babalola for there prayers, to my siblings Mr. Adeleke Babalola, Miss. Itunu Babalola and Mrs. Yetunde Doyin-Ishola (Babalola) for your support in all areas.

Finally to my friends at EMU, Salah, Mclarry, Lekan and Mr Gbenga for your love and support it is well appreciated.

PREFACE

This is a research written in fulfilment of a Master of Science Degree at the department of Mathematics. It involves application of statistics to factual data obtained from the library on average number of students using the library every session. It uses Simple Linear regression to analyze the data, observation and conclusion was made after analysis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

The study of science is a continuous learning process. It aims; relating to details of social or physical phenomenon are specified and checked by assembling, organizing, numerating and analyzing data.

Population: is defined as a set of objects or units that are of interest to study such as bottles filled up in a day a cola drink company.
Sample: is simply some or subset of the actual population such as bottles filled for the first hours at a cola company.

A study of elements in a population is called census, however the study of a census is almost impossible because of time, cost and method/means of measurement. Data's in a population are studied by taking samples from the population and every property observed from the samples is generalized for the population. For samples not reflecting the population are referred to as biased.

Variables are often mention in statistical analysis; it is define or is seen as a characteristics change between objects in a population. They are often regarded as data [2]. They are in two folds; qualitative variables/data and quantitative variables.

Qualitative data are data variable assigned values such as name, colour or labels. They also regarded as categorical variable, while quantitative data's are numeric: They simply are used in measurable quantity such as population of student at the

university. The number of student is a measurable attribute which makes the population a quantitative variable.

Statistics basically deals with Data analysis obtained from a given population or randomly selected samples. It uses some data analysis parameter called descriptive statistics and inferential statistics. According to [3] descriptive statistics is different from inferential statistics. Descriptive statistics describes what the data shows, they are quantitative in nature and are used in a manageable form. It is sometimes called exploratory data analysis, and hence can also be define as investigating variable measurements the in data sets.

These are sometimes relationship between variables or individual variables; Processed data refers to information, or raw data obtained from an experiment or an historical record. If we are analysing students ID cards, for example, descriptive statistics describes the percentage of ID cards issued to students at Eastern Mediterranean University each semester, or the birth certificate issued at the General Hospital Magusa. Any random number chosen for computation are descriptive statistics for the data from which the statistic is computed.

Descriptive statistics gives a full idea of a data. Inferential Statistics on the other hand is concerned with making conclusion about a population from a sample. This is done by random sampling, followed by inferences made about the number of distribution.

A careful study was carried out to determine the average population of students using the library yearly, starting from 2007/2008 session to 2014/2015 session. Histogram chart as already been constructed, to highlight the level of improvement according to

records obtained from the library managements. However, we shall apply a simple regression analysis to this record obtained from the liberian to predict future use of the library, and then draw suggestions and conclusions on it.

Chapter 2

THEORY OF REGRESSION ANALYSIS

In regression analysis, it is seen as a statistical tool for the investigation of relation between variables of at least two datasets. Normally a statistician seeks to ascertain the reasons why a variable differs from one to another. The relationship between a two variables for example could be positive, negative or most time no real relationship exist. All these attribute accounts for the regression index value.

Variable sets are said to be a positive regression if both are increasing at the same time, a negative regression if in a set of variables when one variable is increasing in value while the other is decreasing. In a scenario where there is no trend between the two variables, there is no regression index value. This regression index value is what is called correlation.

2.1 Reason for Regression

As mentioned in previous chapter, regression analysis, evaluates the relationship between variables. It is a method to find a functional relationship between dependent and independent variable. In a situation where claims is often made that crime rate doubles with unemployment. Data experts evaluate past information on this hypothesis to give a definitive prediction on future crime based on current or past information. Researches show benefits of using regression analysis:

1. It reveals, significant relationships between variables, that is the dependent and independent variables; so mathematical relationship is constructed by regression analysis.

2. It shows the strength of impact of the independent variables on the dependent variable.

In analysis of regression, it is possible to compare the effects of measured variables such as the crime rate and unemployment situation. This helps researchers or data scientists to eliminate and evaluate variables to build a model.

2.2 Types of Regression Techniques

The fact remains, that there are various kind of regression techniques for prediction. These techniques are mostly driven by three parameters; Quantity of independent variables, quantity of dependent variables and shape of the regression line. The techniques include;

1. Linear Regression; This is a very common technique used widely. It's a common tool used to determine models, the dependent variable is continuous while independent variables can be continuous or discrete, and the line of regression is linear. Further details on simple regression line will be given in the ensuing pages, however, short definitions will be given for the remaining techniques.
2. Logistic regression; It determines the probability of success and failure in an event. It is only used if dependent variable is binary.
3. Polynomial regression; A regression equation is polynomial in nature if the power of the independent variable is higher than1.
4. Stepwise regression; this form is used when we deal with multiple independent variables.
5. Ridge regression; it is a type of technique used when the data suffers from multi-co linearity.

Simple regression also called linear regression initiate relationship between variable;

a dependent variable and one or more independent variable using best fit straight line called regression line.

Simple linear regression is also viewed in different cases based on the position of the line or by observing the scatter plot. Figure 3.12 [1] detailed the relationship between variables depending on how perfect or non-perfect they are linearly related, while some of the observed scatter plot have no real relationship such as (e) and (f). This relationship can also be referred to as correlation "r" which is an index value that shows the degree of relation between variables. These index values will further be explained with an example in this chapter.

2.3 Idea invention which leads to model invention

The formulation of a problem is often more essential than its solution which may be a matter of mathematical or experimental skill. Albert Einstein [3]. Regression analysis is a statistical tool for discussing the relationship between two variables, where a set of variables (Independent variable) is used to estimate and predict the outcome of the other (dependent variable).

In regression data analysis, there are basic facts needed before developing a fit model to analyse processed data.

Understanding the problem is important, and identifying the types of variables or data you want to work with and its acquisition method. To identify the problem and develop the idea or model correctly, the following must be observed[2]

1. The physical background must be analyzed. Statisticians' often combine effort with others to understand basic things about the subject area, which is an opportunity to learn something new.
2. The aims must be stated, because instances may arise where team members may not understand the impact of what you are doing.

3. It is important to understand what is needed for the problem. Sometimes complicated analyses are performed than needed, where simple descriptive statistics may be all that is required.
4. Always put the problem in statistical terms. This is a challenging step and is error prone if proper caution is not taken. However with the help of some application software and careful imputation we will mitigate the error rate.

Various ways data's are obtained are by experiments and by observations. Data's obtained in this research are based on observation and will discuss further, parameter used to develop it's model.

In regression analysis there are independent and dependent variables which are also called predictor and response variables respectively. The effects dependent variables variables are determined by the effects of their predictor. The goal here is to build a simple model - by using an equation to determine the response for the given independent variables with small error of prediction. The relationship between any two variables can be represented on a *scatter plot*, as will be seen later in the ensuing figures.

A straight line is then drawn using point that produces minimum error, for any X value, exist a variability. The line drawn is the line of means which gives the mean of all values of 'Y' corresponding to a given value of 'X'. Figure 7 is an output from SPSS on a solved question [1], Recall properties of linearity that these figure as a positive a slope, that is relation between change in Y and X and an intercept on Y when X = 0. The straight line intercept sometimes called line of mean is a tool to predict Y for any X value.

Description in previous frame produces hypothesis on which a simple regression analysis is built. They are the following;

1. Average value of dependent variable Y changes by the independent variable X in either increasing or decreasing sense. This infers that relationship between X and Y is linear.
2. Values of Y determined by X are normally distributed whose mean appears on the regression line.
3. There is a noticeable and equal standard deviation for dependent variable Y, for any given value of X variables; that is variability Y with respect to X are the same.
4. There are noticeable deviations of Y variables about the mean, which are not dependent, this deviation of any Y variable with other Y values for any X value is void; that is, there is no effect.

In data analysis involving regression, observations are in two's on any subject. The mean line is an important tool and can be represented [1]; is the regression equation

$$Y=ß_0 + ß_1X + \varepsilon \ldots \text{ eqn (1)}$$

Where '$ß_0$' and '$ß_1$' are the regression coefficients, ε = error.

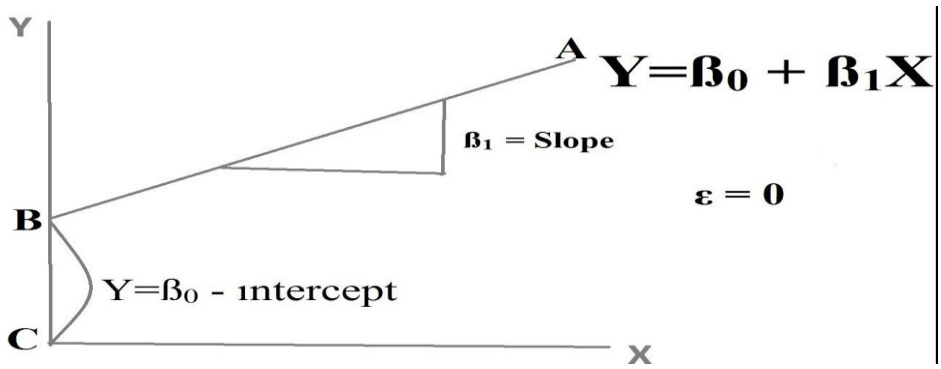

Figure 1: Straight line model

In fıgure 1 shows a fırst order model also called the probabilistic model.
It represents the lıne of means and they are the population parameters with numerical values, thıs values are known when there is access to the entire population of (x,y) measurements; Line A, B is the regression line.

2.4 Fitness model

In figure 6, notice the scatter diagram, a regression line is best fit when distance from line points are low, that is, point lines are required to minimize difference from observed Y for all X values. Points on Y that fall on or close to the line offer better regression line because, it reduces points variation on the line which Akins to reducing errors on line. Any point distance from this line is called fit [1]. (As compared in figure 5 and figure 6) A residual however, is the difference between observed Y, and fitted value. Figure 6 illustrates the residual errors (e).

Regression analysis is used to measure the variability of variables. A simple regression determines how two variables 'Y' and 'X' are related using linear parameters. It is very important in every sphere of life because it is used to predict future occurrences. As in figure 7 is a regression line obtained from data [1] and "simple" means that we are working in two dimensions.

Recall equation (1); Where $ß_0$ = intercept on the response axis

$ß_1$ = slope of the straight line

$\varepsilon = 0$ = error

then; Y= random variable whose value change with X, with errors and X = is an independent variable with negligible error. See Fıg 1.

Hence our regression equation becomes $E(y) = ß_0 + ß_1X$ … eqn (2)

$$Y = E(Y) + \varepsilon \quad \ldots \quad \text{eqn (3)}$$

Equation (2) is a true regression line with data's scattered around the line, for our estimated regression;

$$y = b_0 + b_1 x \quad \ldots \text{ eqn } (4)$$

where b_0 and b_1 are the estimated intercept and slope respectively, and they will be proofed.

y = predicted or fitted value

since our fitted regression line $y = b_0 + b_1 x$, equation (4) becomes the true estimate of equation (2)

from a given regression data (x,y); I = 1,2,…,n

the residual error "e", also called the unexplained error is;

$$e_i = Y_i - y_i \quad \ldots\ldots\ldots \text{ eqn } (5)$$

equation (4) then becomes $e_i = Y_i - y_i = [Y_i - (b_0 + b_1 x_i)]$

for this to be a good fit our i^{th} error must be minimal.

Sum of residual errors "SE" [1] = $\Sigma (Y_i - y_i) = 0$

Sum of the squared errors "SSE" [1] = $\Sigma (Y_i - y_i)^2$

Hence SSE;

$= {}^{n}\Sigma_{i=1} \; e^2_i = {}^{n}\Sigma_{i=1} (Y_i - y_i)^2$

$= {}^{n}\Sigma_{i=1} = {}^{n}\Sigma_{i=1} (Y_i - b_0 - b_1 x_i)^2 \ldots\ldots\ldots$ eqn (6)

Using least square error method,

Differentiate "e" with respect to b_0 and b_1 "de/db_o" in equation (6)

Differentiating (6) with respect to b_0

$$de/db_o = -2 \Sigma (Y_i - b_0 - b_1 x_i) = 0$$

divide both side by 2

$$\Sigma (Y_i - b_0 - b_1 x_i) = 0$$

$$\Sigma Y_i - nb_0 - b_1 \Sigma x_i = 0, \text{ divide both sides by "n" we have;}$$

$$\Sigma Y_i / n - b_0 - b_1 \Sigma X_i / n = 0$$

$$\Sigma Y_i / n - b_1 \Sigma X_i / n = b_0$$

By inspecting the above equation, notice $\Sigma Y_i / n$ and $\Sigma X_i / n$ are averages that is means.

hence $\underline{\underline{Y}} - b_1\underline{\underline{x}} = b_0$ … eqn (7)

which will be our fitted or estimated equation; $\underline{\underline{Y}} = b_0 + b_1\underline{\underline{x}}$

where ; $\underline{\underline{Y}}$ = Predictor or estimator on Y axis with respect to $\underline{\underline{x}}$

$\underline{\underline{x}}$ = independent variables, and b_0 and b_1 are coefficient.

Differentiating **(6)** with respect to b_1

$de/db_1 = -2 \Sigma (Y_i - b_0 - b_1 x_i) x_i = 0$

$\Sigma(Y_i - b_0 - b_1X_i) X_i = 0$

$\Sigma Y_i x_i - b_0\Sigma x_i - b_1X_i^2 = 0$ … eqn (8)

Substitute (7) into (8) we have;

$\Sigma Y_i x_i - (\underline{\underline{Y}} - b_1\underline{\underline{x}}) \Sigma x_i - b_1\Sigma x_i^2 = 0$

$\Sigma Y_i x_i - \underline{\underline{Y}} \Sigma x_i + b_1\underline{\underline{x}} \Sigma x_i - b_1\Sigma x_i^2 = 0$

$\Sigma Y_i x_i - \underline{\underline{Y}} \Sigma x_i = b_1\Sigma x_i^2 - b_1\underline{\underline{x}} \Sigma x_i$

$b_1 (\Sigma x_i^2 - \underline{\underline{x}} \Sigma x_i) = \Sigma Y_i x_i - \underline{\underline{Y}} \Sigma x_i$ … eqn (9)

Multiply both sides by "n"

$b_1 (n\Sigma x_i^2 - n\underline{\underline{x}} \Sigma x_i) = n\Sigma Y_i x_i - n\underline{\underline{Y}} \Sigma x_i$ … eqn (10)

Recall for Mean on "X" and "Y" plane of the linear graph is given as;

$\Sigma x_i / n = \underline{\underline{x}}$ and $\Sigma Y_i / n = \underline{\underline{Y}}$

Therefore; $\Sigma x_i = \underline{\underline{x}}n$ and $\Sigma Y_i = \underline{\underline{Y}}n$

Equation **(10)** becomes;

$b_1 (n\Sigma x_i^2 - \Sigma x_i \Sigma x_i) = n\Sigma Y_i x_i - \Sigma Y_i \Sigma x_i$

Finally b_1; $b_1 = n\Sigma Y_i x_i - \Sigma Y_i \Sigma x_i / (n\Sigma x_i^2 - \Sigma x_i \Sigma x_i)$ … Eqn (11)

We shall demonstrate a regression equation using a question [1].Table 1 [1], presents a live experiment carried out on Sixteen couples on pain empathy and brain activity. Empathy, is define as being able to grasp and abnormally feel what Others are feeling. Neuroscientist at a university examined the correlation between brain activity and pain related empathy on individual who watched others in pain[4]. Sixteen couples participated in the experiment. Painful action were applied to male

partner while female partner (individual) watched, estimates were taken for each female: y =pain related brain activity, x = empathic concern.

Table 1: comparing variable Y and X, using least squares model for prediction

Couple	**Brain Activity (Y)**	**Empathic concern (X)**	**XY**	**X^2**	**X - x`**	**Y-Y` Residual error (e)**	**$(X-x`)^2$**	**$(Y-Y`)^2$ Squared error$(e)^2$**
1	.05	12	0.6	144	-6	-0.2088	36	0.0436
2	-.03	13	-0.39	169	-5	-0.2888	25	0.0834
3	.12	14	1.68	196	-4	-0.1388	16	0.0193
4	.20	16	3.2	256	-2	-0.0588	4	0.003457
5	.35	16	5.6	256	-2	0.0912	4	0.00832
6	0	17	0	289	-1	-0.2588	1	0.06698
7	.26	17	4.42	289	-1	0.0012	1	0.000001
8	.50	18	9	324	0	0.2412	0	0.05818
9	.20	18	3.6	324	0	-0.0588	0	0.00346
10	.21	18	3.78	324	0	-0.0488	0	0.00238
11	.45	19	8.55	361	1	0.1912	1	0.0366
12	.30	20	6	400	2	0.0412	4	0.00169
13	.20	21	4.2	441	3	-0.0588	9	0.00346
14	.22	22	4.84	484	4	-0.0388	16	0.00151
15	.76	23	17.48	529	5	0.5012	25	0.2512
16	.35	24	8.4	576	6	0.0912	36	0.00832

Notice the independent variable, which is our ‘X’ variable are in serial increasing order which is used to predict brain activity. The graph plotted reveals that as

empathy score increases the pain related brain activity also increase with minimal amount of error. We shall apply the equations derived to solve data's in table 1, Use equation… (11) to obtain the slope b_1 of the samples, by inspection the scatter-plot obtained using SPSS on figure 5,

We know it is going to be a positive slope. The parameters of $b_1$are respectively obtained as "XY" and "X^2" in the table.

The summations are; $\Sigma Y = 4.14$, $\Sigma X = 288$, $\Sigma XY = 80.96$,

$\Sigma X^2 = 5362$, $n = 16$

$b_1 = n\Sigma Y_i x_i - \Sigma Y_i \Sigma x_i / (n\Sigma x_i^2 - \Sigma x_i \Sigma x_i)$

$b_1 = 16\ (80.96) – 1192.32 /16(5362) – 82944 = 103.04/2848$

$b_1 = 0.03617$

to obtain b_0 = intercept on Y axis, recall equation……. (7)

then replacing $\underline{Y}$ and $\underline{x}$ with Y` and x` which are respectively means on Y and x axis.

Therefore we have; $Y` = b_1 x` + b_0$

or alternately we use $b_0 = \Sigma Y_i – b_1 \Sigma x_i / n$ [2]

Hence; $Y` = \Sigma Y_i / n = 0.2588$ and $x` = \Sigma x_i / n = 18$,

hence $b_0 = -0.3928$ which is evidence on the graph in figure 6 as an intercept on Y.

Therefore, equation $\underline{Y} = 0.036\underline{x} - 0.39$, is the fitted equation for table 1 as shown in figure 8. Comparing the computed coefficients in the equation to SPSS print out, notice the slope and intercept value are exactly the same figure (2).

2.5 Prediction and Estimation

The modal equation obtained is found satisfying to describe the relationship between brain activity and empathy. Hence, the objective of the question may be changed to determine brain activity if empathy concern is given another value not present in table 1. Interpolation is easily applied using the fitted equation;

$\underline{Y} = 0.036\underline{x} - 0.39$, since our "X" values have just an interval between them, we do not need to interpolate here. Table 2, present the estimated value of "Y" using the fitted equation. Notice, some of the observation compared to their corresponding

estimates are a little different, but for variable fifteen shows a big difference and we could guess it to be an outlier for now.

However, we focus on the extrapolation that is, values not found in the observed variables, but there is limitation to what we can extrapolate using parameters called quartiles, this shall be detailed in section 2.7, but for now let us extrapolate for value 25.45; that is to determine what will be the brain activity of a spouse when painful stimuli of degree 25.45 is applied. Recall the equation; $\underline{\underline{Y}} = 0.036\underline{\underline{x}} - 0.39$, Therefore estimated or predicted "Y" would be 0.526. Table 2 presents the residual error [1] or unexplained error; $Y - \underline{\underline{Y}}$, that is the difference between the observed value and estimated Y value. To obtain these errors we shall recall equation (7) that is the predicted equation, $\underline{\underline{Y}} = 0.036\underline{\underline{x}} - 0.39$, the sum of errors, "SE" obtained from the predicted equation is 0.012 which is approximately zero.

Using $\Sigma(Y - \underline{\underline{Y}})$ … Eqn (13)

That is; $\Sigma(Y - \underline{\underline{Y}}) = 0.012$

Also sum of squared errors "SSE" is $\Sigma(Y - \underline{\underline{Y}})^2$ ……. (14)

Hence using equation (13), SSE is 0.3588.

The calculations involved obtaining b_0 and b_1 and SSE in simple regression are sometimes tasking and prone to errors, even when using a calculator, but fortunately with the improvement in technology and the advents of sophisticated application software such as SAS, MINITAB, MATLAB and SPSS which has made computation quite easy.

However, this research used SPSS as seen in figures 2, 3 and 4, whose output conforms to the manual calculation of table 2.

Table 2: estimated values from predicted equation, and errors comparism

Predicted equation $\underline{Y}$ **=0.036x-0.39**	**Unexplained error** $(Y - \underline{Y}) = e_2$	**Squared error** $(Y - \underline{Y})^2 = (e_2)^2$	**Explained error** $\underline{Y} - Y` = e_1$	**Squared error** $(\underline{Y} - Y`)^2 = (e_1)^2$
0.042	0.008	0.000064	-0.2168	0.047
0.078	-0.108	0.01166	-0.1808	0.0327
0.114	0.006	0.000036	0.1448	0.0209
0.186	0.014	0.000196	-0.0448	0.002
0.186	0.164	0.0269	-0.0728	0.0053
0.222	-0.222	0.0493	-0.0368	0.0014
0.222	0.038	0.0014	-0.0368	0.0014
0.258	0.242	0.0586	-0.0008	0.00000064
0.258	-0.058	0.0034	-0.0008	0.00000064
0.258	-0.048	0.0023	-0.0008	0.00000064
0.294	0.156	0.0243	0.0352	0.0012
0.33	-0.03	0.0009	0.0712	0.0051
0.366	-0.166	0.0276	0.1072	0.0115
0.402	-0.182	0.0331	0.1432	0.0205
0.438	0.322	0.1037	0.1792	0.0321
0.474	-0.124	0.0154	0.2152	0.0463

Model	Unstandardized Coefficients		Standardized Coefficients	T	Sig.
	B	Std. Error	Beta		
(Constant)	-.392	.220		-1.787	.096
VAR00002	.036	.012	.627	3.015	.009

Figure 2: Portion of SPSS Print out on regression Coefficients

Then we obtain the Coefficient of Determination R^2 [1]; Coefficient of determination is an index value that helps to predict how well variables are related, that is it explain the linearity of X and Y.

Formula for R^2 is given [2];

$R = \{\Sigma XY - 1/n\ (\Sigma X)(\Sigma Y)\} / (n-1)\delta_x\delta_y$ … Eqn (12)

Where δ_x and δ_y are respectively deviations on X and Y axis.

$\delta_x^2 = \Sigma(X - x`)^2 /n - 1 = 178/15$, $\delta_x = 3.445$

$\delta_y^2 = \Sigma(Y - Y`)^2 /n-1 = 0.59186/15$, $\delta_y = 0.1986$,

inserting values into equation(12), we have:

$R = 80.96 - 0.0625(1192.32) / 15\ (0.6417) = 0.6275$

$R^2 = 0.3939$ approx 0.394, which is same as obtained using SPSS.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.627[a]	.394	.350	.16008

Figure 3: Portion of SPSS print out for Empathetic concern

The interpretation for R^2 means about 39% of the sample variation on brain activity can be explained by using empathy concern "X" to predict brain activity "Y" with least square line.

Model	Sum of Squares	Df	Mean Square	F	Sig.
Regression	.233	1	.233	9.092	.009[b]
Residual	.359	14	.026		
Total	.592	15			

Figure 4: Portion of SPSS print out for errors

Notice the obtained SSE in equation (14), which is also seen in Table 2 column 3, we shall call it unexplained error "e_2" which is also the same as equation (13). "SEE" which is sum of squares of explained errors as seen in table 1, column 4; given by $\underline{\underline{Y}} - Y` = e_1$, is the difference between the calculated or estimated Y and the mean on Y axis. Hence total error is; $e = e_1 + e_2$

Logically, total error is the difference between our observed variables in question and the mean on Y axis, as seen in Table 1, column 7. Recall equation (12) which was used to obtain coefficient of determination "R", we shall use the explained error and the total error to obtain same coefficient R and compare our answers.

$R^2 = \Sigma e_1^2 / \Sigma e_2^2 = (\underline{Y} - Y`)^2 / (Y - \underline{Y})^2$ … eqn (15)

summation of column 5 in table 1 = $\Sigma e_1^2 = 0.2274$

summation of column 3, table 2 = Σe_2^2

therefore $R^2 = 0.384$, R = 0.619. This answer is nearly the same as the SPSS print out in figure 4, and the calculated value in eqn..(12)

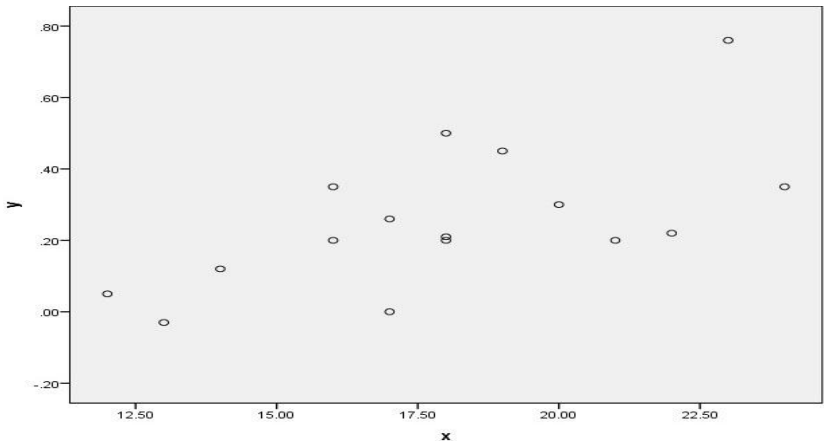

Figure 5: SPSS print out of pain empathy and brain activity

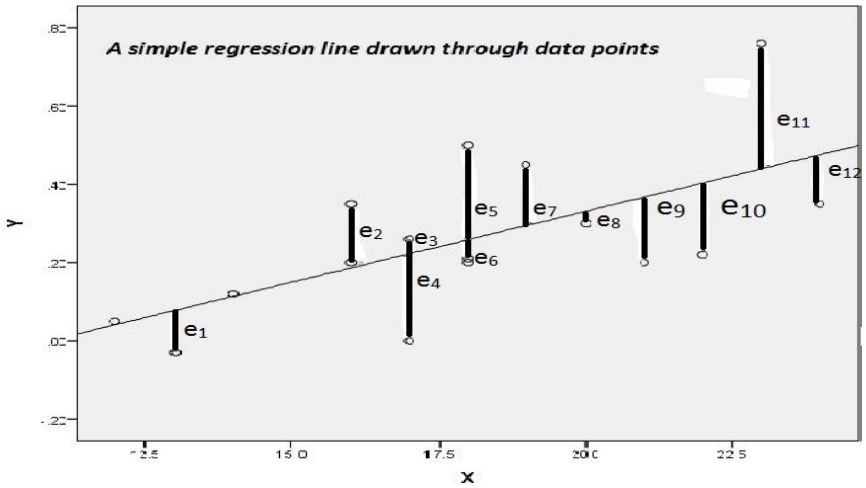

Figure 6: individual variable error around the true mean

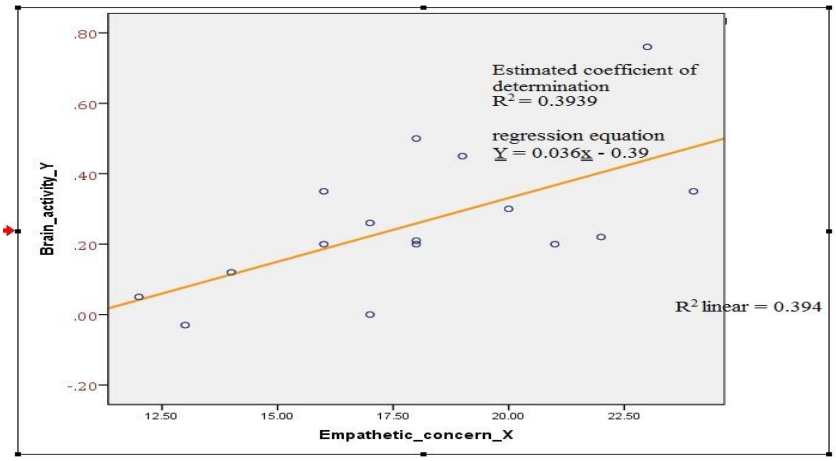

Figure 7: regression line of brain activity and empathy concern

2.6 Outliers

Outliers are observations that fall far from the clustered points in a scatter plot, they are points that are not best fit for a regression coefficient. They occur often in factual data and are inconspicuous, this because, in present century (21^{st}) large data are analyzed by computer and may lead to improper inspection. Outlier in most case observed cases are results of typing errors and misplaced decimals. Looking at figure 6, we could descriptively say point with error e_{11} is an outlier. Some of this outlier tend to be influential to the least square-line. These influential points could have a good or bad effect on the least line [10]. A good influential point is usually found in the X axis, this are points removed from the bulk point and then found reasonably close to the regression line, a bad influential point on the other hand are placed, distance away from the least line, around which major points are clustered, they sometimes called regression outlier [9].

There are various method proposed by researches to compute for outliers [6], [9], [10], but this research used SPSS for evaluations and so manual computation are based on parameters used by the software. These parameters are called quartiles: To define quartiles, we introduce three variables Q_1, Q_2, and Q_3. Q_1 is called the lower quartile, values below which 25% of the data lies. Q_3 is the upper quartile,

with values below which 75% of the data lies and Q_2 is the median of the data set. Quartiles split into four in ordered data sets. To obtain quartiles, we consider Empathic concern (X variables) in Table 1;

$Q_2 = ½ (X_{n/2} + X_{n/2+1})$ for even sets of variables … eqn (15)

$Q_2 = (X_{n+1} / 2)$ for odd sets of variable … eqn (16)

In equations (15) and (16), X_n is the variable position in the whole data set.

$Q_1 = ½ (X_{n/2} + X_{n/2+1})$ for even and $(X_{n+1} / 2)$ for odd … eqn (17)

Where "n" is the total number of variables *upto* the median Q_2.

$Q_3 = ½ (X_{n/2} + X_{n/2+1})$ for even and $(X_{n+1} / 2)$ for odd … eqn (18)

Where "n" is the total number of variables from the median Q_2.

Inter-quartile range = $Q_d = Q_3 - Q_1$ … Eqn (19)

Extreme values are; $X_i < Q_1 – 1.5Q_d$ and

$X_i > Q_3 + 1.5Q_d$ … eqn (20)

Therefore, from equation (15), (17) and (18) our quartiles are respectively $Q_2 = 18$, $Q_1 = 16$ and $Q_3 = 20.5$ then equation (19) and (20) becomes; $Q_d = 4.5$ and extreme values are 9.25 and 27.25.

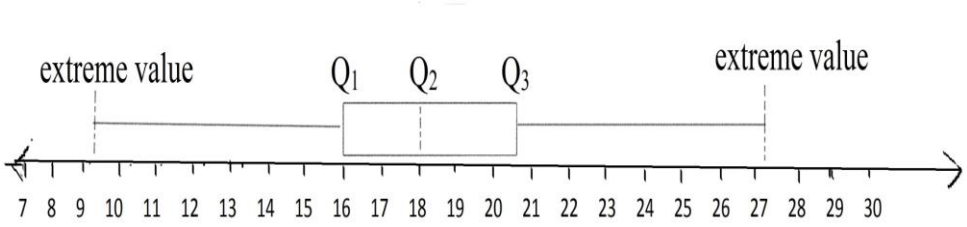

Figure 8: box plot on Empathic concern.

In table 1, variable X does not have an outlier from the box plot in figure 5 contrary to earlier statement made which is by inspecting the scatter plot. Prediction equation $\underline{\underline{Y}} = 0.036\underline{x} - 0.39$ is only valid for variables within 9.25 through to 27.25, that is variables 10, 11, 25, 26 and 27 which is not in table 2 can have its estimated or predicted value.

However variables $X_i \leq 9$ and $X_i \geq 28$ is not valid with the prediction equation obtained from equation (7).

Chapter 3

DATA ANALYSIS

In this chapter, we shall apply the method with equations derived in previous chapter to analyze data's obtained from the library. These data's are records of students on campus that makes use of the library every sessions.

The information was obtained through a clocking machine which all students must past through with an ID cards; this was done so to properly ascertain the quantity of students entering the library with no errors. Although some student come with no real intention, that is some come to play, sleep, talk to friends' e.t.c. However majority students are there to study as the record as shown over the years.

3.1 Design Method for SPSS

SPSS is one of many applications used to analyze data, there are lots of information obtained from it; it can display graphical shapes of data, including histogram, pie chart. SPSS package used here is a version 20 created by IBM. It can be obtained from the University's laboratory, once it is installed; it creates a short-cut on the desktop for easy access. Double click to launch, it as graphical interface which makes it a little easy to use. As shown in figure 9, by the left bottom corner is the data view and the variable view.

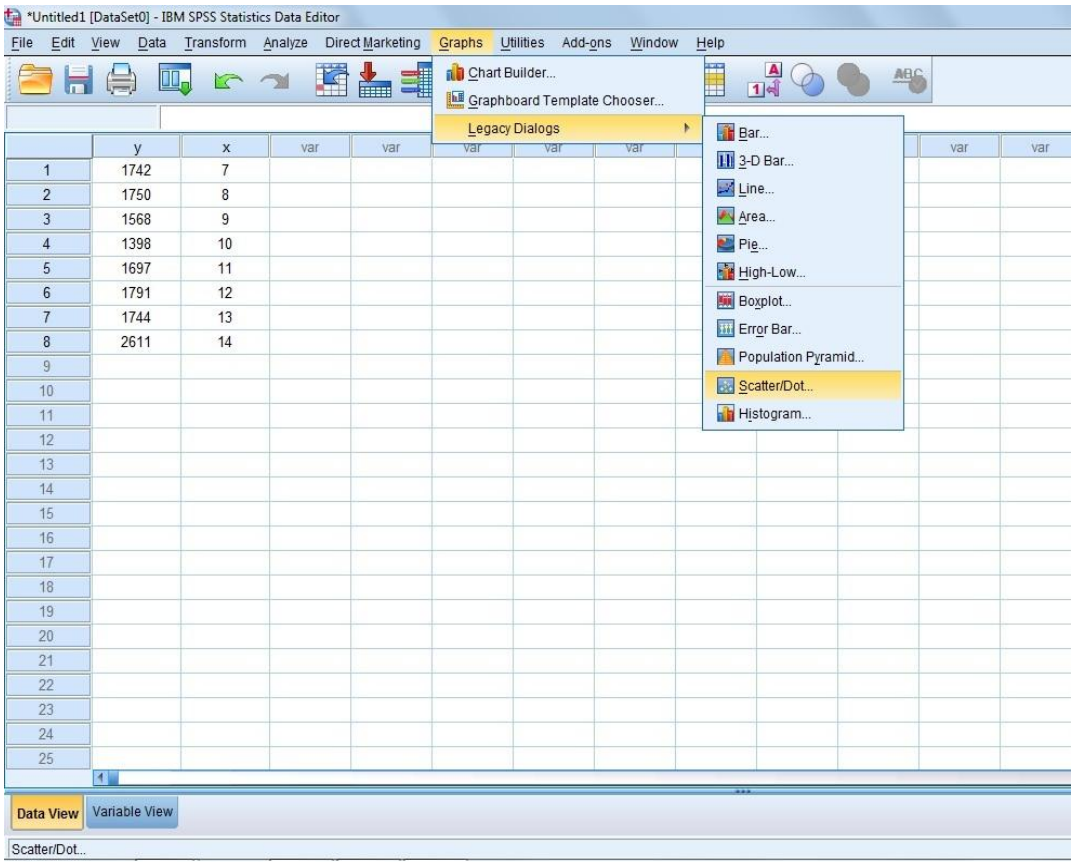

Figure 9: SPSS interface

Data view: it is simply an interface where data are inputted into the working environment as shown above. Variable view: it as a lot of parameters which affect how data's on the data view are placed, this parameters includes type of data such as numeric, comma, date and currency, decimal place of figures. Figure 14 is a scatter plot with the regression line obtained by navigating through the tool bars as in figure 9 Graph > legacy dialogue > scatter plot. Figures 11 – 13 are obtained by navigating through the tool bar in figure 9; Analyze > regression > linear, this will bring a dialogue box as shown in figure 10, then match variables X and Y to independent and dependent variable column respectively and press the ok button.

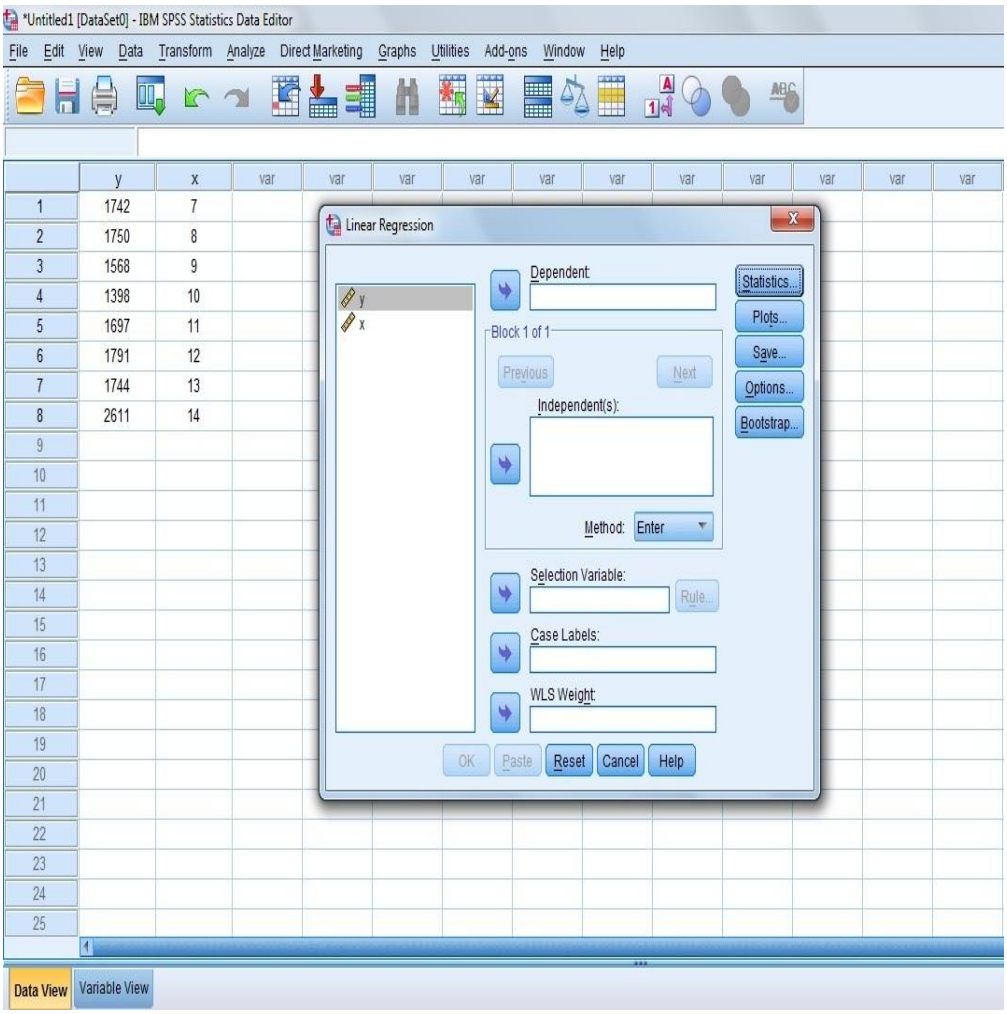

Figure 10: SPSS interface

Table 3: shows data's obtained from the library director from year 2007 to year 2014

Average population (Y)	1742	1750	1568	1393	1697	1791	1744	2611
Year since 2007 (X)	7	8	9	10	11	12	13	14
Year/session	2007 /2008	2008 /2009	2009 /2010	2010 /2011	2011 /2012	2012 /2013	2013 /2014	2014 /2015

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.573[a]	.328	.216	316.67403

Figure 11: Coefficient of regression

Figure 14, from elementary graph plotting shows a positive slope and hence produces a positive correlation "R" as seen in column 2 in figure 11; this means, as values of "Y" increases, values of "X" also increases partially. Coefficient of determination "R^2" which is 0.328 shows that approximately 33% sample variation in the library usage can be explained using yearly population data.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	293837.357	1	293837.357	2.930	.138[b]
	Residual	601694.643	6	100282.440		
	Total	895532.000	7			

Figure 12: Errors generated

The standard error of the estimate in column 5 of figure 11 is obtained from the residual error. Figure 14 is the ANOVA output of SPSS; it is the analysis of variance and degrees of freedom "df". Sum of squares of unexplained errors in row 2 is obtained using equation (14) and dividing by 6. Six which is the degree of freedom means 6 variable point are responsible for the errors obtained, that is total samples used are 8 but 2 plots will give a perfect positive correlation, and as plot begins to increase degrees of freedom tends to increase and "R" and "R^2" coefficients begin to diminish. Mean square of residual is obtained by dividing equation (14) by "n-2"

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	908.750	525.144		1.730	.134
x	83.643	48.864	.573	1.712	.138

Figure 13: Regression coefficient

Figure 13 is the output of coefficient values of regression line.

Regression equation is written as $\underline{\underline{Y}} = 908.750 + 83.643\underline{\underline{x}}$….. (21)

The prediction equation is also useful to an estimate library usage for the past four sessions and for the future four sessions, any prediction outsides this range makes the prediction void. The quartiles parameters are used to make prediction on library usage for the coming sessions, equations 17-20 is used to obtained the quartiles hence, $Q_1 = 8.5$, $Q_2 = 10.5$, $Q_3 = 12.5$ and $Q_d = 4$.

It extreme values are 2.5 and 18.5, which means equation (21) can be used to predict library usage up to 2018 and it can also give an estimate of population in 2005 prior to 2007 when data collection started. Using equation (21), gives an average estimate of 2163 students that will use the library in 2015/2016 session; also estimates for 2016/2017 session and 2017/2018 session are respectively 2247 and 2331.

3.2 Scatter plots (influential points)

Looking at figure (14), points "σ" and "Δ" are outlier with, point "Δ" having positive influence on the least square line and point "σ" having a bad influence [10]. Notice also $R^2 = 0.328$ shows that approximately 33% sample of variation in library usage can be explained by using yearly population data given.

Figure 13 however, is an SPSS output when point "σ" which we assume to have a

bad influence on the least line is absent. Notice a better "R^{2}" which shows that 36% of sample can be explained. R^2 obtained shows a weak regression because it is less than half of one, hence our equation maybe a best fit model, although a non-linear model will be best fit on variables obtained from library.

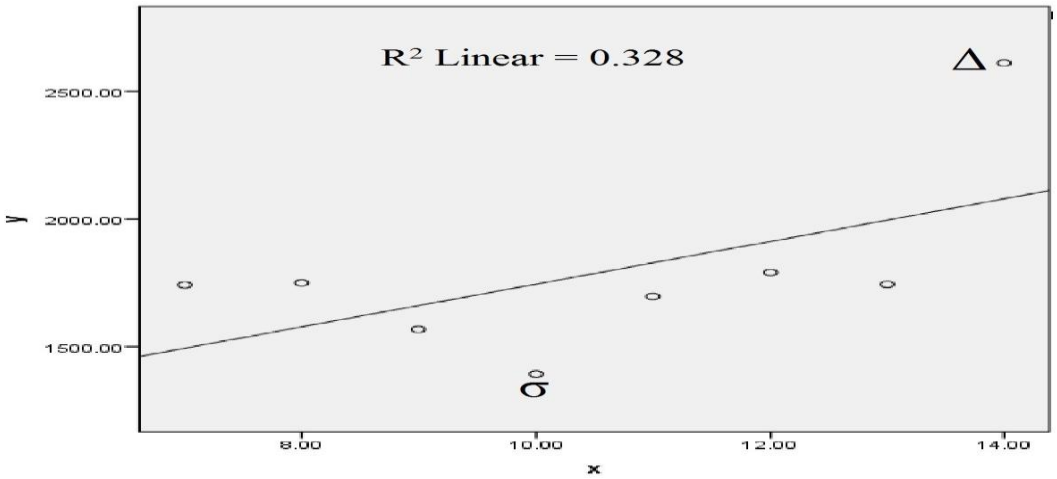

Figure 14: SPSS output on data obtained from library

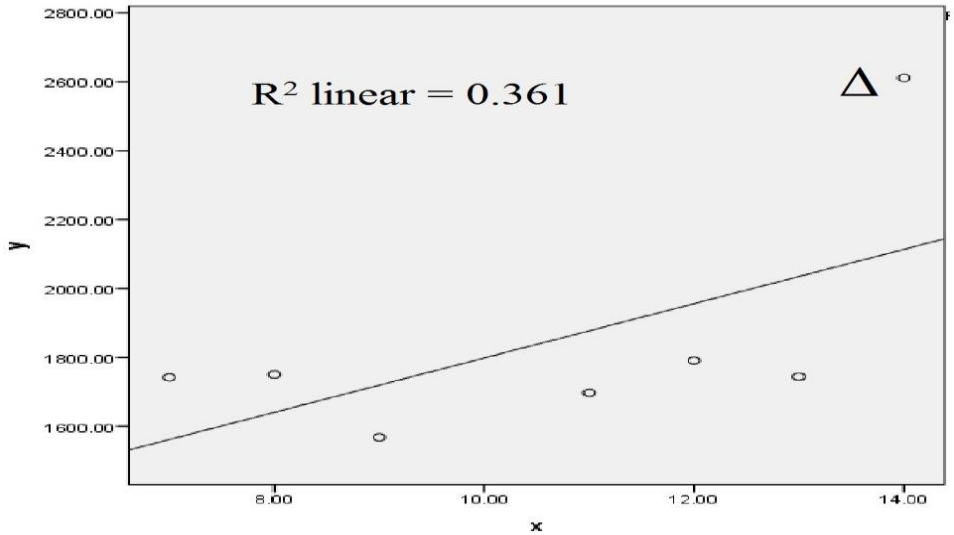

Figure 15: SPSS output without point "σ"

3.3 Skewness of the variables

To determine skewness;

Skewness = (mean - median) / standard deviation … eqn (22)

If a positive value is obtained it is positively skewed, and negative skewed if a negative value. A zero value means the distribution is normal. Therefore from table 3, independent variable "X" has mean and median values to be 10.5.This makes the numerator of equation (22) to be zero after the difference, then zero divided by anything will always be zero.

This proves that the "X" variables are normally distributed. Y variables follow same principle, where the median and mean are respectively 1743 and 1787; then divide their differences with the standard deviation which gives a positive value, hence Y variables is positively skewed.

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
X	8	7	14	10.50	2.449	6.000	0.000	0.752
Y	8	1393	2611	1787.00	357.677	127933.14	2.027	0.752

Figure 16: SPSS output of Descriptive Statistics

Chapter 4

DISCUSSIONS AND OBSERVATIONS

Simple regression is an important tool in applied statistics, but it remains a pedagogical neglect and controversy. However in light of data analysis from the library R^2 obtained shows approximately 33% of the data can be explained and hence makes the model somewhat a best fit. The least line equation in Equation..(21), can be used to predict future average population of students that will use the library for the next three session; respective sessions of 2015/2016, 2016/2017 and 2017/2018 shows an estimate of 2163, 2247 and 2331, and no radical increase for each session but a gradual increase.

In figure14, point "σ" does have a bad influence on the least line. Figure 15 gives a better R^2 value when the bad influence point was removed. Point "Δ" in figure 14 however is an outlier but it as a positive influence on the least line. Data analysis in this research is best analyze using a non-linear model.

REFERENCES

[1] William, M., Terry, S., (2011). *A course in Statistics regression analysis. Seventh edition.*

[2] Julian, J., Faraway (July 2002). *Practical Regression and Anova using R.*

[3] William, M. K., Trochim, (2008). *Research method knowledge based*.

[4] Singer, T., (February 20, 2004). Science Journal retrieved from http://www.sciencemag.org (page 6)

[5] Warren, M., *Regression with SAS*. University of California. Retireved from http://www.ats.ucla.edu/stat/sas

[6] Richard, W., (January22, 2015) Outliers review. Retrieved from http://www3.nd.edu/~rwilliam/

[7] Wilcox. R:R: *Fundamentals of modern statistical methods Springer*, New York 2001, ISBN 0-387-95157-1

[8] Dallal, G. E., Regression diagnostics. Retrieved from http://www.tufts.edu/gdallal/

[9] Olive. D,. *Applied robust statistics*. Preprint M-02-006, Retrieved from http:www.math.siu.edu/

[10] Rousseeuw, PJ-Leroy, A. M: *Robust regression and outlier detection*. J .wiley, New jersey 2003. ISBN 0-471-48855-0.