A Statistical Analysis on the Visits to EMU Health Center by the Students

Ogheneovo Mclarry Eduiyovwiri

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

Masters of Science
in
Applied Mathematics And Computer Science

Eastern Mediterranean University
February, 2016
Gazimağusa, North Cyprus, Turkey

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Cem Tanova
Acting Director

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

Prof. Dr. Nazim Mahmudov
Chair, Department of Mathematics

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

Examination Commitee

1. Prof. Dr. Sonuç Zorlu
2. Asst. Prof. Dr. Rashad Aliyev
3. Asst. Prof. Dr. Mehmet Ali Tut

ABSTRACT

There are different statistical techniques in estimating and predicting future events or outcome given a set of independent factors influencing such an event. Regression analysis is one of the modern statistical tools used for such purpose. Knowing the outcome of an event given a set of independent variable will help make proper decisions regarding the scenario. Here, regression analysis was used to predict the number of visitors visiting some key department of the Eastern Mediterranean University Health Center. This will help the school management to know the area where the health center is shorting man power and to also carry out a research or study on the reason why visitors are faced with such illness relating to the department they visit often. A solution has been detected and discussed to help in the prediction of the number of visitors visiting some key department of the school health center. A regression analysis has been carried out on the data set of the visitors who visited the health center in the past 22 months (January, 2014 to October, 2015) this involves the number of visitors in each month and the department they visited. This is done by the use of statistical software called SPSS. It is use for regression, and prediction measure especially when one is dealing with large numbers.

Keywords: Estimating, Department, Health Center, Predicting, Regression Analysis, SPSS, Statistical Techniques.

ÖZ

Gerçek hayat olaylarında (uygulamalarında) bilinmeyen (var olmayan) parametre değerlerini kestirimini yapabilmek için istatiksel metodlardan Regresyon analizi önemli bir rol oynamaktadır.Bağımsız parametre değeri kullanılarak bilinmeyen değer bulunan regresyon fonksiyonu yardımıyla bulunabilmektedir.

Yapılan bu çalışmada DAÜ Sağlık merkezine başvuran hastaların hangi ünite(branş) üzerinde yoğunlaştıkları Eregrasyon analizi yardımıyla modellenerek gelecek aylarda beklenen ziyaretçi sayıları kestirilmesiyle çalışılmıştır. Yapılan kestirimlerde üniversitesinin yoğunluk yaşayacağı söylenebilir.

Anahtar kelimeler : Bağımsız değişken , Bağımlı değişken , Kestirme, Öngörme ,regresyon analizi ,SPSS

To God Almighty and My Lovely Family

ACKNOWLEDGEMENT

I give thanks and praise to Jehoval God for his love, mercy and grace for making the thesis work a reaity, I also appreciate my beloved parents and family for their undiluted support toward the actualiazation of this work. May God Bless you all.Words will not be enough to say thank you to my late beloved uncle, Brother Ochuko Esievwerhie whom before his death was very supporting, may God bless your soul. I wont fail to appreciate the humble assistance of my supervisor, Asst. Prof. Dr. Mehmet Ali Tut with whom advice and encouragement made this thesis a dream come true, may God bless you sir. Lastly I will say a very big thank you to my friends here and back homefor their love and support during my study, God bless you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Statistical science is the prcedure of gathering, organizing, sorting out, examing and interpretation of data (numerically) with the end goal of settling on a solid choice. When talking of arrangements even remotely with the collection, handling, translation and presentation of information(numerical) fits in with the space of statistics.From the definition above, it shows clearly that there are steps or stages in statistical science.

First is the collection of data of interest such as the amount of rainfall in a year, the number of students admitted by the institute of graduate studies in past 10 academic session, the scores of students in a particular MTH test and so on. Having collected our data of interest, its good we organize them in a way or manner where we can/could be able to carry out analysis process. Analyzing the organized collected data of interest is crucial for the purpose of carrying out the statistical process in the first place. Information like the range amount of rainfall in a year, the average performance of students in a MTH exam , the session in which the institute of graduate studies admitted more Africans students than other nationalities.

Interpreting the analyzed data is one of the most import part of a statistical science because as a statistician it’s expected that you interpret whatever numerical result you get from analyzing a set of data in a real life word in a way that seems

meaningful to lane man. For instance a statistician should be able to explain what he meant when he said or says the average amount of rainfall in a particular place in a year is 6 per year and so on.

Finally, making estimation and prediction is the end product or reason for carrying out a statistical process in the first place. After a careful and comprehensive statistical analysis, we should be able to make estimation and prediction on some future events thereby making us to taking appropriate decisions. For instance after a careful analysis on the amount of rainfall and we predict that the amount of rainfall for the coming year, will be higher than the present, then an umbrella producing company can make a decision on producing more umbrella and this will yield more profit for the umbrella producing company who has gotten a prior knowledge on the prediction on the amount of rainfall in the coming year than those who never had such knowledge.

Researchers do follow the bloodline of Statistics to 1663 when the record on Natural and Political Observation upon the Bills of Death Rate by John Grout was published [1] while Statistical Science (Lovric, 2000) has shown that there has been an increase in the statistical techniques and methods from the late 1930. In the early 20^{th} century. Francis Galton and Karl Pearson changed statistics into a vital thorough arithmetic field of study utilized for investigation as a part of science as well as in commercial ventures, government and other circle of life [2].

Today there is no field or branch of life that does not need the application of statistical process. From health to commerce to politics, you name it. The invention of todays sophisticated computers and statistical software has made carrying out

statistical process a much easier task as it helps carrying out statistical processes on a large number of data

1.1 Data Analysis

Statistics as verbally expressed earlier has to do with the analyzing of data into more abstract information, this is because when we carry out an analysis on a set of data, it gives us, an incipient set of data which avails us to understanding the information contained in the experimented data. The objective of statistics is to expand understanding from data, for this the understanding the basic concepts associated with data analysis becomes important.

In applying statistics to a real life problem, the first thing we do as explained earlier is the collection of data of interest which statistically is term the population and sample. Populace can be described as the arrangement of individual persons or objects in which an examiner is essentially fascinated amid his or her examination situation that is population is the total or entire data associated with the case to be studied. And this can be a huge task as testing every single man in the country of interest will not be an easy task, for this, taking a sample out of the population come to play. Sample is a subset of the entire population. For example, carrying out a experiment on the effect of a new drug in tackling malaria on men in a country, the population here is the total number of men in the said country of interest. The task of testing every single man in the country which will certainly be a killing task, we can decide to test a sample of men that is testing some group of men randomly from all corners of the country, by so doing we have reduced the number of men to be counted. Statistical Science can be descriptive when we are dealing with organizing, displaying and explaining a given sets or data or it could be Inferential when it

involves drawing a conclusion and prediction of future events based on the result gotten from analyzing a sample from the population of interest. Descriptive and inferential statistics are intewoven as one can say inferential statistics is the utilization of the the results gotten from descriptive statistice in making solid decisions Analyst often desire to know how much effect a variable or more has in the determination of an outcome. To this end, the study of regression analysis has been a very important tool for statisticians in analyzing the relationship between two or more variables and as well as estimating and predicting future values for given sets of independent variables.

Regression analysis a statistical tool used in modelling a mathematical function to describing the relationship between two variables, the independent and the dependent variable. It avails in presaging a future replication from a given independent variable utilizing a mathematical function which avails in reducing the error of presage and this is possible through the utilization of the least square method. This method was first utilized in the 1805 by Legendre [3] and Gauss in 1804[4]. Both Legendre and Gauss utilized the least square method in determining from astronomical experiments, the orbits f bodies about the sun and since then, regression has been a consequential implement for analyzing, estimating and predicting data.

In the 1990s, economists made utilization of the least square method in estimating economics events. Today with the avail of the advanced computers and statistical software, carry out a regression analysis has been a fascinating and less arduous process as statisticians can carry out the regression process on immensely colossal scales.

This thesis will intend to apply the least square regression method in modelling a mathematical function which will be utilized in discussing the relationship between the mean numbers of total visitors visiting the university health center here in Eastern Mediterranean University with key department of the health center. That is we optate to ken if there is a relationship between the total number of visitors that visits the health center and the number of these visitors visiting key department, this will help know the kind of medical threatment they mostly go for, is it internal medicine, skin, ocular perceiver or is it bone quandary they come for most. And we shall be utilizing the modelled mathematical function in soothsaying the number of students whom will be visiting the health Centre in the future. Erudition of these will avail the school and the Health Centre management to take decision on which department in the Health Centre that needs more man power or equipment as well finding denotes to study the cause of such illness here on campus.

In the next chapter, we shall be talking and discussing the various theory in regression analysis as well as how to estimate and predict a future value or response given a set of variables with the use of a regression model. We shall also be discussing the errors in predicting and how to minimize the error of predicting and estimating as well as the confidence and prediction intervals.

Chapter 2

THEORY OF REGRESSION ANALYSIS

Regression analysis is a statistical tool for discussing the relationship between two variables where one (the independent variable) is used to estimate and predict the outcome or response of the other (dependent variable).

In practical when one or a statistician is called upon to discuss the relationship between variables and asked the outcome of a certain event given previous data of the variables, if given the information on the number of hours students spent in preparing for an exam with their corresponding grade in the exam. Here the number of hours the student spent in preparing is the independent variable and the grade of the students are the outcome or response or the dependent variable. A useful means or model used in showing the relationship between the outcome y and the independent variable x is

$$y = \beta_0 + \beta_1 x \qquad (2.1)$$

with β_0 and β_1 being the $y-$ intercept and slope respectively, the relationship is shown graphically below

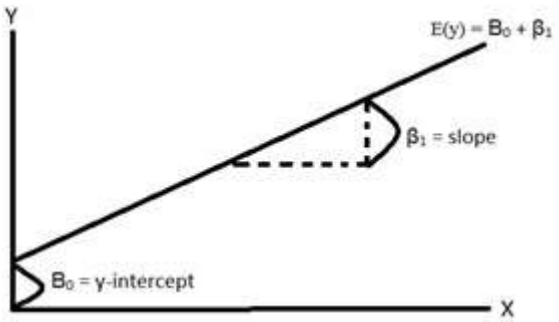

Figure 1: The Straight line model

The idea of regression analysis deals with considering the best relationship between y and x, measuring the strength of the relationship and using methods that permits for prediction of the response values y given values of the regressor x.

2.1 The Fitted Regression Line

A significate side of the regression analysis is to calculate the β paramenters , that is calculating the values of the regression coefficients. This is usually done by the leat square method. The calculated or fitted regression line is given as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad (2.2)$$

Where $\hat{y}_i$ is the i^{th} predicted value of y for every change in x

$\hat{\beta}_0$ is the y intercept of the regression line

$\hat{\beta}_1$ is the slope of the regression line

Apparently, the fitted line is used as an appraisal of the real regression line and we envision that the fitted line should be closer to the real regression line when a boastfully number of data is useable.

2.2 The Least Square Method

If we are talking of regression analysis, then we are talking about the least squares which is all about reducing the error of estimating and predicting to nearest

minimum. The process of minimizing the parameter estimates of SSE is what is called the least square.

Mathematically;

$$SSE = \sum_{i=1}^{N} e_i^2 \qquad (2.3)$$

We want to find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the measure in the above equation; this is executed by differential calculus. In ascertaining the minimum value of a function using calculus is simply differentiating the said function and equating the derivative to zero. Thus, if we want to find the value of $\hat{\beta}_0$ and $\hat{\beta}_1$that minimizesSSE, we need to express SSE in terms of $\hat{\beta}_0$ and $\hat{\beta}_1$, and differentiates SSE with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$. And equate each to zero and solve for $\hat{\beta}_0$ and $\hat{\beta}_1$.

However, since SSE consist of two important paramenters, $\hat{\beta}_0$ and $\hat{\beta}_1$, we will find the partial derivatives of SSE with respect to each of the parameters while treating the other invariable.

$$\text{SSE} = \sum_{\text{i=1}}^{\text{N}} \text{e}_\text{I}^2 = \sum_{\text{i=1}}^{\text{N}} (\text{y}_\text{i} - \hat{\text{y}}_\text{i})^2$$

$$= \sum_{\text{i=1}}^{\text{N}} \left(\text{y}_\text{i} - \left(\hat{\beta}_0 + \hat{\beta}_1 \text{x}_\text{i}\right)\right)^2 = \sum_{\text{i=1}}^{\text{N}} \left(\text{y}_\text{i} - \hat{\beta}_0 - \hat{\beta}_1 \text{x}_\text{i}\right)^2$$

differentiate SSE w. r. t $\hat{\beta}_0$

$$\frac{\partial \text{SSE}}{\partial \hat{\beta}_0} = \frac{\partial}{\partial \hat{\beta}_0}\left[\sum (\text{y}_\text{i} - \hat{\beta}_0 - \hat{\beta}_1 \text{x}_\text{i})^2\right] = \sum \left[\frac{\partial}{\partial \hat{\beta}_0} (\text{y}_\text{i} - \hat{\beta}_0 - \hat{\beta}_1 \text{x}_\text{i})^2\right]$$

We treat y_i, $\hat{\beta}_1$ and x_i as constants

$$\frac{\partial SSE}{\partial \hat{\beta}_0} = \sum [-2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)]$$

$$= -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \quad (2.4)$$

Differentiating SSE w.r.t $\hat{\beta}_1$

$$\frac{\partial SSE}{\partial \hat{\beta}_1} = \frac{\partial}{\partial \hat{\beta}_1} \left[\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \right] = \sum \left[\frac{\partial}{\partial \hat{\beta}_1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \right]$$

We treat y_i, $\hat{\beta}_0$ and x_i as constants

$$\frac{\partial SSE}{\partial \hat{\beta}_0} = \sum [-2x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)]$$

$$= -2 \sum x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \quad (2.5)$$

Equate (2.4) to zero and multiply by $-{}^{1}/_{2}$, we have

$$\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \Rightarrow 0 = \sum y_i - \sum \hat{\beta}_0 - \sum \hat{\beta}_1 x_i$$

make $\sum \hat{\beta}_0$ the subject

$$\sum \hat{\beta}_0 = \sum y_i - \sum \hat{\beta}_1 x_i$$

Since $\hat{\beta}_0$ and $\hat{\beta}_1$, are the same for all instances in the original equation, this farther simplifies to

$$N\hat{\alpha} = \sum_{i=1}^{N} y_i - \hat{\beta}_1 \sum_{i=1}^{N} x_i$$

$$\hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{N}$$

$$\hat{\beta}_0 = \frac{\sum y_i}{N} - \hat{\beta}_1 \frac{\sum x_i}{N}$$

(2.6)

It can be seen that $\frac{\sum y_i}{N}$ is the mean of y_i while $\frac{\sum x_i}{N}$ is the mean of x_i, therefore

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (2.7)$$

Now we move back to equation (2.5) for $\hat{\beta}_1$; multiply both sides by $-1/2$, we have

$$0 = \sum x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum (y_i x_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2)$$

$$0 = \sum y_i x_i - \sum \hat{\beta}_0 x_i - \sum \hat{\beta}_1 x_i^2$$

$$0 = \sum y_i x_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2$$

$$\hat{\beta}_1 \sum x_i^2 = \sum y_i x_i - \hat{\beta}_0 \sum x_i$$

$$\hat{\beta}_1 \sum x_i^2 = \sum y_i x_i - \left[\left[\frac{\sum y_i}{N}\right] - \hat{\beta}_1 \left[\frac{\sum x_i}{N}\right]\right] \sum x_i$$

$$\hat{\beta}_1 \sum x_i^2 = \sum y_i x_i - \frac{\sum y_i \sum x_i}{N} - \hat{\beta}_1 \left[\frac{(\sum x_i)^2}{N}\right] \text{ collect like terms}$$

$$\hat{\beta}_1 \sum x_i^2 + \hat{\beta}_1 \left[\frac{(\sum x_i)^2}{N}\right] = \sum y_i x_i - \frac{\sum y_i \sum x_i}{N}$$

$$\hat{\beta}_1 \left(\sum x_i^2 + \frac{(\sum x_i)^2}{N}\right) = \sum y_i x_i - \frac{\sum y_i \sum x_i}{N}$$

$$\hat{\beta}_1 = \frac{\sum y_i x_i - \frac{\sum y_i \sum x_i}{N}}{\sum x_i^2 + \frac{(\sum x_i)^2}{N}} \qquad (2.8)$$

A second partial differentiation of SSEwith respect to each paramente, we have from (2.4)

$$\frac{\partial \mathrm{SSE}}{\partial \hat{\beta}_0} = -2\sum \mathrm{x_i}\left(\mathrm{y_i} - \hat{\beta}_0 - \hat{\beta}_1 \mathrm{x_i}\right)$$

$\frac{\partial^2 \mathrm{SSE}}{\partial \hat{\beta}_0{}^2}\left[-2\sum \mathrm{x_i}\left(\mathrm{y_i} - \hat{\beta}_0 - \hat{\beta}_1 \mathrm{x_i}\right)\right]$ treating $\mathrm{y_i}.\hat{\beta}_1$ and x_i as constants, we have

$$-2\left(-\sum\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)\right) = 2\sum\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) \qquad (2.9)$$

Also finding the second partial derivatives of SSE with respect to $\hat{\beta}_1$ we have from (2.5)

$$\frac{\partial SSE}{\partial \hat{\beta}_1} = -2\sum x_i\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)$$

$$\frac{\partial^2 SSE}{\partial \hat{\beta}_1^2} = -2\left[\sum x_i(-x_i)\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)\right] = 2{x_i}^2\sum\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) \qquad (2.10)$$

Since both second partial derivatives are non-negative, we can be sure that the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that quantify the equations negated by setting each partial derivative to zero are the minimum values of SSE.

Another way of representing a linear regression model is the matrix form given below

Given a data set $\{y_i, x_{i1}, x_{i2}, \ldots, x_{ik}\}_{i=1}^{n}$ of n units, linear regression model assumes the relationship between the independent variable y_i and k-vector of predictor x_i. Observations recorded for each of the n level can be expressed in the following way;

$$y_i = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \varepsilon_2$$

.

.

$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \varepsilon_n$$

And this n-system of equations can be expressed follows;

$$y = X\beta + \varepsilon \tag{2.11}$$

Where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ . \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & x_{1n} \\ 1 & x_{21} & x_{22} & . & . & x_{2n} \\ \vdots & \vdots & \vdots & \cdot & \cdot & \cdot \\ 1 & x_{n1} & x_{n2} & . & . & x_{nn} \end{bmatrix} \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ . \\ \beta_n \end{bmatrix} \text{ and } \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ . \\ \varepsilon_n \end{bmatrix}$$

The matrix X is known as the design matrix, it carries the information about the levels of independent variables at which the observation are obtained. The vector β contains all the regression coefficients relating to the independent variables. To form or create a regression model, β should be known and its estimated using the least square estimates. The equation below is used ;

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{2.12}$$

Knowing the estimates $\hat{\beta}$, the linear regression model can then be estimated by

$$\hat{y} = X\hat{\beta} \tag{2.13}$$

And the error of estimate ε is the given as

$$\varepsilon = y_i - \hat{y}_i$$

2.3 Simple Linear Regression

By “Simple” we mean that we are dealing with a two dimensional surface as in the case of a flat piece of paper. It doesn’t mean the theory here is easy.

The simplest graphical model for relating a response y for every individual independent variable x is drawing of a straight line through a plotted data points of the y-response against the independent variable x. For this reason, Simple linear regression can also be referred to as a straight line regression model.

When we have a bunch of points on a scatter plot, we can draw a line that seems to represent a general trend, such a line is called regression line. This is different from the true line, mathematically; the regression line is given by

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Where $\hat{y}$ is the predicted value, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the y-intersect and the slope respectively, and x is the value to be estimated or predicted.

The difference between the true line and the regression line is simply the difference in the errors and the residuals. .

Understanding their differences is very important in regression theory when predicting.

Algebraically;

True line; $y = \beta_0 + \beta_1 x$

Regression line; $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

The $\frown$ on the variable shows that these are not the true numbers or variables but rather they are estimated variables. Therefore the regression probabilistic equation is given as;

$\widehat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x + U_i$ (2.3.1)[4]

Where $\widehat{y_i}$ is the i^{th} predicted value of y for every change in x

$\hat{\beta}_0$ is the y intercept of the regression line

$\hat{\beta}_1$ is the slope of the regression line

U_i are the residuals which is the distance from the i^{th} predicted point to where it touches the regression line.

The best method for drawing a regression line is the least square method, which means finding the line that best minimizes the sum of the squares of the residuals. As stated earlier, the regression line is given as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Where

$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$ (the slope of the least square regression line)

Where

$\bar{x}$ is the mean value of the independent variable

$\bar{y}$ is mean value of the dependent variable

And

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ (the intercept of the least square regression line)

2.3.1 Model Assumption

It will be good to revisit the simple linear regression probabilistic model presented earlier

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \varepsilon$$

and discuss graphically how the true line is. Lets turn our attention to the random error ε, we want to see how the random error determines how best the model define

the genuine association between the response and the independent variables, from the graph below, it can be seen that the points plotted are (x, y) points scattered along the true line.

It can be seen that each line point is a normal distribution of its own to the center of the distribution falling on the line.

The mean of the errors ε over an limitless long arrangement of a process is zero for each independent variable x. that is

$$E(y_i - \hat{y}_i) = 0. \qquad (2.3.2)[5]$$

It can also be seen that all distribution of ε have the same variance say σ^2.

And the distance between each individual y to the point on the line will be its individual ε value that is the error for each point is unique.

Since,

$$y_i - E(y_i) = y_i - (\beta_0 + \beta_1 x_i) = \varepsilon_i$$

Therefore, for any (x, y), the associated deviation ε all have variance σ^2.

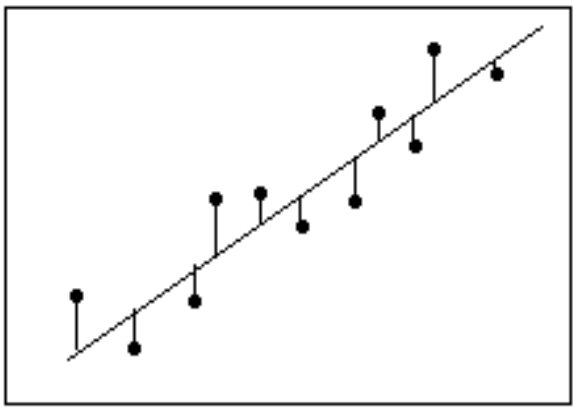

Figure 2: Individual observation around the true line

2.3.2 Variance of Estimators

Drawing an inference in $\beta_0 \, and \, \beta_1$, it is important we arise at an estimate of the parameterσ^2. σ^2 is the model error variance or experimental error variation around the true regression line. For clarification, let's use the notation;

$$S_{xx} = \sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{2.3.2}$$

$$S_{yy} = \sum_{i=1}^{N}(y_i - \bar{y})^2 \tag{2.3.3}$$

$$S_{xy} = \sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \tag{2.3.4}$$

We write the error sum of squares as;

$$SSE = \sum_{i=1}^{N}(y_i - \hat{y})^2 \qquad \left(\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x\right)$$

$$= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum \left[(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})\right]^2$$

$$= \sum (y_i - \bar{y})^2 - 2\hat{\beta}_1 \sum (x_i - \bar{x})(y_i - \bar{y}) + \beta_1^2 \sum (x_i - \bar{x})^2$$

$$= S_{yy} - 2\hat{\beta}_1 S_{xy} + \hat{\beta}_1^2 S_{xx}$$

$$= S_{yy} - \hat{\beta}_1 S_{xy}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{2.3.5}$$

It seems fair to assume that the bigger the deviation of the random error ε (calculated by its variance σ^2), the bigger error in the estimation of the model $\hat{\beta}_0$ and $\hat{\beta}_1$ will be[5].

Furthermore, $\bar{y}_i$ is used in estimating in the latter sample situation, while $\hat{y}_i$ is used in estimating the mean of y_i in a regression structure.

s^2 is an unbiased estimator of σ^2 and the $(n-2)$ divisor which is the degree of freedom associated with s^2. In standard normal distribution $(n-1)$ divisor which is of one degree of freedom is subtracted from n, the cause is that one parameter is estimated which is the mean μ by $\bar{y}$ but in regression, two parameters are estimated which are $\beta_0 \, and \, \beta_1$ by $\hat{\beta}_0 \, and \, \hat{\beta}_1$ respectively. Therefore the parameter σ^2 is estimated by

$$s^2 = \frac{SSE}{n-2} = \sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)^2}{n-2}$$

and we refer to s as the mean square error[5] .

Its adviceable we compute to six significant figure the values of $SS_{yy}, \hat{\beta}_i's, and \; SS_{xy}$ when executing SSE operation as we might effect our model.[5].

It is worth to note that about 95% of the perceptions exist in 2s of their separate least square predicted value $\hat{y}$[5].

2.3.3 Hypothesis Testing on the Slope (β)

An important t-test on the slope is the hypothesis test that

$$H_0: \beta = 0 \quad \text{vs} \quad H_1: \beta \neq 0$$

If the null hypothesis (H_0) is accepted, the entailment is that the mean E(y), does not change as x changes[10]. In a simple regression model, it means the true slope β_1 is equal to 0 as shown in the figure below and we conclude that there is no enough proof to demonstrate that x contributes data to the determination of y. On the other hand, if the data supports the alternative hypothesis then we reject the null

hypothesis(H_0), and we conclude that there is a significant relationship between E(y) and the independent variable x.

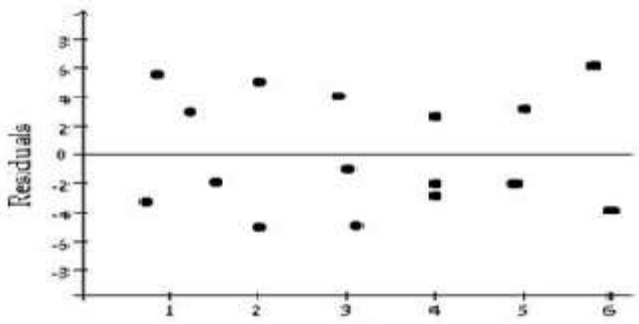

Figure 3: Model with slope equals zero

Its worthy to note that we use t-test because the population parameters here with which were working with does not entails the entire population but rather a sample of the population is being worked with. And this is the reason we used the sample variance s^2 in estimating the population variance σ^2.

2.3.4 Sampling Distribution of $\widehat{\boldsymbol{\beta}}_1$

In the event of ε, $\hat{\beta}_1$, the least squares of estimator of the slope will be a normal distribution with mean β_1 (the true slope) and the standard deviation

$$\sigma_{\hat{\beta}} = \frac{\sigma}{\sqrt{SS_{xx}}} \qquad (2.3.7)$$

but since population variance σ will usually be unknown, the appropriate test statistic will generally be the use of a sample variance *s*,where

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}} \qquad (2.3.8)$$

Test of model; Simple linear Regression

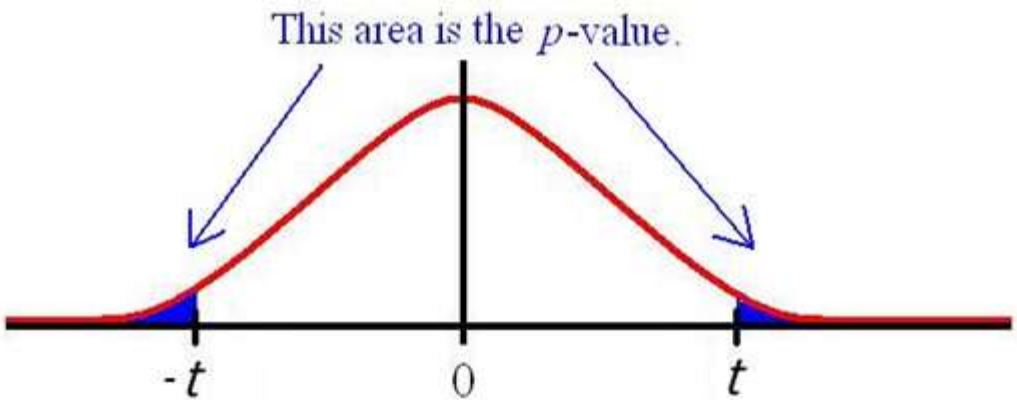

Figure4: T-Distribution

Mathematically, t-test statistics is computed as

$$\mathrm{t} = \frac{\hat{\beta}_1 - \text{Hypothesized value of } \beta_1}{\text{Standard eror}} = \frac{\hat{\beta}_1}{\mathrm{s}/\sqrt{\mathrm{SS_{xx}}}}$$

with $(n-2)$ degree of freedom to establish a critical region.

T-Test statistics for Simple Linear Regression

Test statistics $\mathrm{t} = \frac{\hat{\beta}_1}{\mathrm{s}/\sqrt{\mathrm{SS_{xx}}}}$

One-Tailed Test

	$\mathrm{H_0}: \beta_1 = 0$	$\mathrm{H_0}: \beta_1 = 0$
	vs	
	$\mathrm{H_1}: \beta_1 < 0$	$\mathrm{H_1}: \beta_1 > 0$
Rejection region	$\mathrm{t} < -\mathrm{t}_\alpha$	$\mathrm{t} > \mathrm{t}_\alpha$
p − value	$P(t < t_c)$	$P(t > t_c)$

Two-Tailed Test

$$H_0: \beta_1 = 0 \text{ vs } H_0: \beta_1 \neq 0$$

Rejection region	$\lvert t \rvert > t_{\alpha/2}$
$p-$value	$2P(t > t_c)$ if t_c is positive
	$2P(t < t_c)$ if t_c is negative

A $100(1-\alpha)\%$ confidence interval for the slope β_1 in a simple linear regression is

$$\hat{\beta}_1 - t_{\alpha/2}\frac{s}{\sqrt{SS_{xx}}} < \beta_1 < \hat{\beta}_1 + t_{\alpha/2}\frac{s}{\sqrt{SS_{xx}}} \qquad (2.3.9)$$

where $t_{\alpha/2}$ is a value of the t-distribution with $(n-2)$ degree freedom.

2.3.5 The Coefficient of Correlation

Coefficient of correlation otherwise known as the Pearson Moment coefficient of correlation have to do with the measure of strength of the linear relationship between two variables x and y. The basic idea of correlation is to report if there is an association between the x and y, it helps us to know if there is a positive, negative or no relationship between the independent variable and the dependent variable and its computed as follows;

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{xy}}} \qquad (2.3.10)$$

Values of r is always between -1 and +1, a value of r close to 0 means lack of relationship between x and y. On the other hand, a value of r close to -1 or +1 shows a negative or a positive correlation respectively. A negative correlation implies the higher the value of the independent variable the lower the dependent variable and a positive correlation means the higher the independent variable the higher the dependent variable. A rare occasion is the point when $r = 1$ that is a perfect correlation.

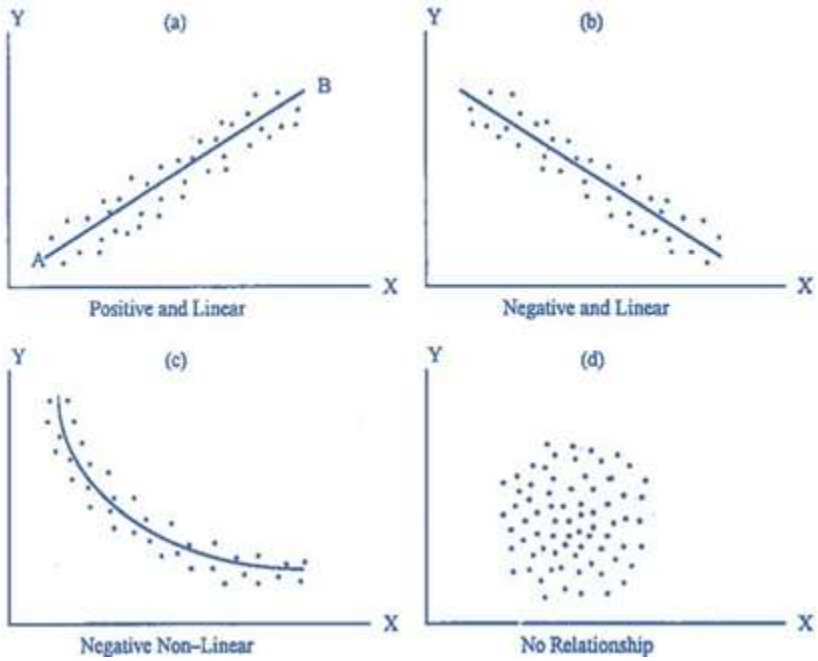

Figure 4 Values of r and their implicatios

2.3.6 Coefficient of Determination (r^2)

One of the ways of measuring the utility of a regression model is to quantify the contribution of x in predicting the response y. it is the proportion of the total variation in the dependent variable y that is explained or accounted for by the variation in the independent variable x. it's a convenient way of measuring how well the least squares equation perform as a predictor of y is to compute the reduction in the sum of the square of deviations that can be attributed to x, expressed as a proportion of SS_{yy}.

Mathematically, coefficient of determination is computed as;

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{Explained\ sample\ variability}{Total\ sample\ variability} = 1 - \frac{SSE}{SS_{yy}}$$

Where $SS_{yy} = \sum(y_i - \bar{y})^2 \ \ and \ \ SSE = \sum(y_i - \hat{y}_i)^2$ is the sum of the least squares of deviation. Its worthy to note that in a regression model, if there is a little or no

association between x and y then both $SS_{yy}\ and\ SSE$ will be nearly equal and in such case $r^2 will\ be\ equal\ 0$. However, if x contributes to the determination of y then $SSE\ will\ be\ smaller\ than\ SS_{yy}$ and if all points of the scatter plots falls on the regression line then $SSE = 0$.

2.3.7 Model Utilization

When a helpful model has been achieved in showing the assiociation of a given independent variable and the corresponding depend variable, then we are ready to accomplishing the primary aim of this study which is estimating and predicting,

Estimating and predicting are the two most common use of a probabilistic model , we use the least square model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

to both estimate the mean value of y for a specific value of x and to predict a particular value of y for a given value of x.

The standard error of estimate is used to establish confidence intervals when the sample size is large and the scatter around the regression line approximates the normal distribution. The more we deviate away from the mean of the independent variable, the larger our error or variation will be and we need to adjust this.

The standard error of the estimator $\hat{y}$ of the mean value of y at a particular value x, say x_a is

$$\sigma_{\hat{y}} = \sigma\sqrt{\frac{1}{n} + \frac{(x_a - \bar{x})^2}{SS_{xx}}} \qquad (2.3.11)$$

Where σ is the standard deviation of the random error ε and we refer to $\sigma_{\hat{y}}$ as the standard error of $\hat{y}$.

The standard deviation of the prediction error for the predictor $\hat{y}$ of an individual y-value for $x = x_a\ is$

$$\sigma_{(y-\hat{y})} = \sigma\sqrt{1 + \frac{1}{n} + \frac{(x_a - \bar{x})^2}{SS_{xx}}} \qquad (2.3.12)$$

Where σ is the standard deviation of the random errorε. We refer to $\sigma_{(y-\hat{y})}$ as the standard error of prediction.

As stated earlier, we are interested producing an interval estimate of two type, confidence interval which report the mean value of y for a given x. And prediction interval which reports the range of values for y for a particular value of x.

A $100(1-\alpha)\%$ Confidence Interval for the Mean Value of y for $x = x_a$ is given as

$$\hat{y} - (t_{\alpha/2})s\sqrt{\frac{1}{n} + \frac{(x_a - \bar{x})^2}{SS_{xx}}} < y < \hat{y} + (t_{\alpha/2})s\sqrt{\frac{1}{n} + \frac{(x_a - \bar{x})^2}{SS_{xx}}} \qquad (2.3.13)$$

Where $t_{\alpha/2}$ is based on $(n-2)$degree of freedom.

A $100(1-\alpha)\%$ Prediction Interval for a Particular $y\ for\ x = x_a$ is given as

$$\hat{y} - (t_{\alpha/2})s\sqrt{1 + \frac{1}{n} + \frac{(x_a - \bar{x})^2}{SS_{xx}}} < y < \hat{y} + (t_{\alpha/2})s\sqrt{1 + \frac{1}{n} + \frac{(x_a - \bar{x})^2}{SS_{xx}}} \qquad (2.3.14)$$

Where $t_{\alpha/2}$ is based on $(n-2)$degree of freedom.

We ought to observe that when we utilize the least square equation comparison to evaluate the mean estimation of y or to anticipate a specific estimation of y given an estimate x value that is beyound the range of the x values given in our information

might prompt blunders of estimation and predictions that are more bigger than anticipated.

2.4 Multiple Regression Analysis Models

The word "multiple" as used in this contest refers to having or consisting of more than one elements ,so in this contest it means a statistical tool used in examine the relationships between two or more independent variables to a dependent variable. Most practical applications of regression analysis consist of more than one factor(independent variables) which is influencing the determination or predicting an outcome of a dependent variable unlike the the straight-line or simple linear regression that deals with just an independent variable.

In this case, multiple regression analysis helps us estimate and predict the value/outcome of an independent variable y given $x_1, x_2, \ldots\ldots\ldots\ldots, x_k$ independents variables. For example, in predicting a rice yield per acre(y), it will depend on quality of seedx_1, soil fertility(x_2). Amount of rain fall(x_3). favorable temperature(x_4) and quality of seeds(x_5).

2.4.1 Types of Multiple Regression Analysis

Generally, multiple regression models is of the form;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \varepsilon \qquad (2.4.1)$$

Where y is the dependent variable, $x_i's$ are the independent variables, β_0 is the y-intercept and $\beta_i's$ $i = 1,2,3,,k$ determines the contribution of individual independent variable x_i and are called the regression coefficients.

The above kind of multiple regression model is known as the first order regression model, it is first order because all our independent variablesx_i's are raised to the power of 1.

However, x_i may be of higher order terms for quantitative prediction e.g $x_2 = {x_1}^2$ and this allows for curvature in the relationship and not a straight line as in the earlier case. This form of model is called a second-order model or quadratic model and mathematically

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \qquad (2.4.2)$$

However this work is focusing on the first kind which is the First-Order multiple regression analysis.

2.4.2 Steps ın Analyzing a Multiple Regression Model

Analyzing a module which has two or independent variables is similar to that of simple regression module where we have just a single variable with some little additional procedures. Below are the steps in analyzing a multiple regression model

1. Random data samples collection (ie the collecting the values of independent variables $(x_i's)$) for each experimental unit of the sample.
2. Model Hypothesis; this involves selecting the independent variables to include in the model.
3. Estimate the unknown parameters (ie the regression coefficients $(\beta_i's)$) using the least square method.
4. Specify the probability distribution of the random error component ε and estimate its variance σ^2.
5. Statistically discuss the utility of the model.
6. Check to see the assumption on the standard deviation σ are satisfied and if mandatory , modify the model
7. If the model is seen fit and accurate, use it in estimating $E(y)$ or predict y for a given value of $x_i's$.

2.4.3 Model Assumption

From equation (2.4.1), it follows the $\beta_i's$ and the $x_i's$ are not random and hence they are determinitics but ε is random for it is independent or unique for every single case. Therefore y is made up of both a deterministic part and a random part. We should note y is random for its value is determined by the actions or contribution of a set of determined independent variables.

$$y = \underbrace{\beta_1 + \beta_1 x_1 + \beta_2 x_2 + \cdots . + x_k x_k}_{\text{determimistic part of the model}} + \underbrace{\varepsilon}_{\text{random error}}$$

As in simple linear model, ε can be positive or negative for any given $x_1, x_2, \ldots, x_k$ values. The mean value and variance of ε are 0 and σ^2 respectively.Matthematically,

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k$$

2.4.4 A First-Order Model with Quantitative Predictors

A first-order model does not include a higher power other than one. That is all independent variables are raised to the power of one.

A first-order model with 6 quantitative independent variable is given as;

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$ where $x_1, x_2, \ldots, x_6$ are the six independent variable determining the outcome of y and β_0 is the value of y when all six variable are zero and $\beta_1, \beta_2, \ldots, \beta_6$ are the coefficient of the $x_i's$ which are the mean change iny for a one limit change in $x_i's$.

2.4.5 Fitting the Model

The Least Square Method is also used in fitting a multiple regression model.

Recall from simple regression model, a least square model is given or estimated by;

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k \qquad (2.4.3)$$

And it minimizes the sum of squared residuals

$$\text{SSE} = \sum(\text{y}_\text{i} - \hat{\text{y}}_\text{i})^2$$

As in the case of simple linear regression model, the value of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, .., \hat{\beta}_\text{k}$ can be obtain from an n-set of simultaneous linear equation. The only differences between this and that of the simple linear regression model is the computational difficulties as the $(\text{n} + 1)$ simultaneous equation which must be solved to obtained $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, .., \hat{\beta}_\text{k}$ are often tedious and time consuming, but with invention of statistical software one can carry out the operations with ease.

2.4.6 Estimating the Variance σ^2 and the Variance of ε

Remember that σ^2 is the variance of the random error ε which is the measure of deviation from the predicted $\hat{y}$ from the true y value. And as such, σ^2 is an important tool when measuring model utility. It shows that if $\sigma^2 = 0$, then the random error $(\varepsilon) = 0$ and consequently $\hat{y} = y$. That is having a perfect prediction where al predicted values of $\hat{y}$ are same as the true values of y. On the other hand, the larger the deviation (the value of σ^2), the greater the error in estimating the model $\beta_0, \beta_1, \beta_2, \dots, \beta_\text{k}$ coefficients and same as the distance between y and $\hat{y}$. For this reason, σ^2 plays an important role in making inferences about $\beta_0, \beta_1, \beta_2, \dots, \beta_\text{k}$ in the utility of the model. We ought to use the result of the regression analysis to estimate the value of the variance σ^2 as it is rarely known.

The mean value of the squares of the distance of dependent variables for a given sets of $\text{x}_1, \text{x}_2, \dots, \text{x}_\text{k}$ values about the mean value E(y) is σ^2 and since the predicted value $\hat{y}$ measures the mean values of y for each x_i's, it makes sense to use

$$SSE = \sum(y_i - \hat{y}_i)^2$$

to formulate an estimate for σ^2.

For a multiple regression model with k independent variable, σ^2 is estimated as

$$s^2 = \text{MSE} = \frac{\text{SSE}}{\text{n}-\text{number ofestimated parameters}} = \frac{\text{SSE}}{\text{n}-(\text{k}+1)} \qquad (2.4.7)$$

Therefore s(standard deviation) is given as;

$$s = \sqrt{s^2}$$

A useful interpretation or meaning of s is that the interval $\pm 2s$ will provide a rough estimate to the accuracy with which the model will predict future values of y for a given sets of x_i's values.

In multiple regression model, we must have to estimate the $(k+1)$ parameters$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, .., \hat{\beta}_k$, hence the estimation of σ^2 is SSE divided by n-number of measured components.

2.4.7 Utility Testing

In simple linear regression model, we demonstrated how to conduct a t-test on β where

$$H_0: \beta_i = 0 \text{ against } H_1: \beta_i \neq 0$$

However, in multiple regression model where a large number of parameters are involved, we need a global test when testing the utility of a multiple regression model, that is for multiple regression model;we test

$$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

VS

$$H_1 \text{ At least one of the } \beta_i' \text{sis non} - \text{zero}$$

The test statistics for this kind of situation is the F-statistics.

$$F = \frac{(SS_{yy} - SSE)/k}{SSE/[n-(k+1)]} = \frac{Mean\ Square(model)}{MSE(\ mean\ Square\ for\ error)} \qquad (2.4.8)$$

We reject H_0(null hypothesis) when $F > F_\alpha$ where F is the value of the above equation with the numerator having a k-degree of freedom while the denominator having $n-(k+1)$degree of freedom and n is the sample size, k is the number of terms in model.

Rejection of H_0 takes place when $\alpha > p - value$, when $p - value$ is given as $P(F > F_c)$, F_c is the calculated value of the test statistic given by the formula above.

The global F-test is regarded as the test, the model must pass to merit its further consideration so when the null hypothesis H_0 in the global F-test is rejected, it does not necessarily mean the model in question is best but rather we say its statistically useful with a $100(1-\alpha)\%$ confidence for another model can be created proving more useful in terms of predicting and estimating.

Inferences about the Individual β parameters as in the case of simple linear regression model. However, this is limited to the parameters the analyst seem important for predicting the value of y to avoid too many type 1 errors which is rejecting a null hypothesis when actually it's true.

So, individual test of β is given as follows;

One-Tailed Test

$$H_0: \beta_i = 0 \qquad H_0: \beta_i = 0$$

$$vs$$

$$H_1: \beta_i < 0 \qquad H_1: \beta_i > 0$$

$$Test\ statistics\ t = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$$

Two-Tailed Test

$$H_0: \beta_i = 0\ \ vs\ \ H_0: \beta_i \neq 0$$

Reject $H_0\ when\ \ t < -t_\alpha\ ,\ \ t > t_\alpha\ and\ \ |t| > t_{\alpha/2}$

where $t_\alpha\ and\ t_{\alpha/2}$ are based on $n - (k + 1)$ degrees of freedom.

n is the number of observations

$k + 1$ is the number of β parameters in the model.

2.4.8 Multiple Coefficient of Determination $R^2\ and\ R_a^2$

R^2 tells us how well a multiple regression model fits a set of data. Like in straight line regression model, its values are between $-1\ to + 1$ which means at $R^2 = 0$ implies a lack of fit of the model. On the other hand at $R^2 = \pm 1$ implies a perfect fit of the model. The multiple coefficient of determination $\boldsymbol{R^2}$ is gotten or calculated by the equation below;

$$R^2 = 1 - \frac{SSE}{SS_{yy}} \quad 0 \leq R^2 \leq 1 \qquad (2.4.9)$$

$$where\ SSE = \sum (y - \hat{y})^2\ and\ SS_{yy} = \sum (y - \bar{y})^2$$

A substitute to coefficient of determination R^2 is the adjusted multiple coefficient of determination denoted by R_a^2. Both have the same readings, however, the later put into consediration the sample size n and the number of β components in the model and its value are or will always be less than that of R^2 even if we add more n samples. Hence its values can never be equal 1.

Mathematically R_a^2 is given as ;

$$R_a^2 = 1 - \left[\frac{n-1}{n-(k+1)}\right]\left[\frac{SSE}{SS_{yy}}\right] \quad (2.4.10)$$

$$= 1 - \left[\frac{n-1}{n-(k+1)}\right](1 - R^2)$$

$$R_a^2 \leq R^2 \quad (2.4.11)$$

$R_a^2 \; and \; R^2$ are just sample statistics and so shouldn't conclude that a model is perfect or not from the values obtained from them. It is advisable for analyst to make use of F-test for testing the global utility and once a model has be deemed useful using the F-test overall utility, $R_a^2 \; or \; R^2$ is used to further measure the variation ratio of y explained by the model.

2.4.9 Utility of the Model

Like in simple linear regression model, we used the least square model in estimating $E(y)$ and predicting the value of y when $\hat{\beta}_0 \; and \; \hat{\beta}_1$ has been calculated for a given x value which must not be far above the maximum data value or far below the minimum data value. Luckily, both $E(y) \; and \; y$ values are same. This is done by replacing the value of $x \; say \; x_a$ into the model

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_a$ and calculate the value of $\hat{y}$.

Same approach is used here in multiple regression analysis, we use the multiple regression model or function to estimate $E(y)$ and predict the value of y for a given set of $x_i's$ values by simply replacing the values of $x_i, i = 0,1,2,..,k$ into the model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

$where \; \hat{\beta}_i's$ has been calculated then we solve for $\hat{y}$.

2.5 Using SPSS for Regression Analysis

When dealing with a multiple regression analysis that is a problem with more than one variable or a simple regression with a large amount of observations, a manual calculation of our regression coefficients will be a very difficult task and we will be open to computational errors and most times we might be tempted to rounding up our figures though it advisable we round up to six significant figures.

However, with the invention of statistical software one of which is SPSS has made it a lot easier for scientist to statisticians to carry out regression analysis with ease. In this work, we shall be giving the steps in using SPSS in carrying out regression analysis and also show how to interpret our outcomes.

2.6 Worked Example

The British Journal of Sports Medicine (April 2000) published a study of the effect of massage on boxing performance.Two variables measured on the boxers were blood lactate concentration(mM) and the boxer's perceived recovery(28-point scale).Based on information provided in the article, the data in the table below were obtained for 16 five-round boxing performnces, where a massage was given to the boxer between rounds.Conduct a test to determine whether blood lactate level(y) is linearly related to perceived recovery(x). Use

Table 1: 16 five-round boxing performance

Blood lactate level	Perceived recovery
3.8	7
4.2	7
4.8	11
4.1	12
5.0	12
5.3	12
4.2	13
2.4	17
3.7	17
5.3	17

5.8	18
6.0	18
5.9	21
6.3	21
5.5	20
6.5	24

2.6.1 Solution

First we calculate the values of $\hat{\beta}_0 \; and \; \hat{\beta}_1$ which is as follows

Table 2: Computations of Regression Parameters

x_i	y_i	$x_i y_i$	x_i^2
7	3.8	26.6	49
7	4.2	29.4	49
11	4.8	52.8	121
12	4.1	49.2	144
12	5	60	144
12	5.3	63.6	144
13	4.2	54.6	169
17	2.4	40.8	289
17	3.7	62.9	289
17	5.3	90.1	289
18	5.8	104.4	324
18	6	108	324
21	5.9	123.9	441
21	6.3	132.3	441
20	5.5	110	400
24	6.5	156	576
$\sum x_i = 247$	$\sum y_i = 78.8$	$\sum x_i y_i = 1264.$	$\sum x_i^2 = 4193$
$\bar{x} = \frac{\sum x_i}{n} = \frac{247}{16}$ $= 15.375$	$\bar{y} = \frac{\sum y_i}{n} = \frac{78.8}{16}$ $= 4.925$		

$$SS_{xx} = \sum x_i^2 - n(\bar{x})^2 = 4193 - 16(15.4375)^2 = 379.9375$$

$$SS_{xy} = \sum x_i y_i - n(\bar{x}\bar{y}) = 1264.6 - 16(15.4375 \times 4.925) = 48.125$$

Therefore the slope $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{48.125}{379.9375} = 0.126666$

And the y-intercept $\hat{\beta}_0 = \bar{y} - \hat{\beta}_i\bar{x} = 4.925 - (0.126666 \times 15.4375) = 2.969594$

There the regression equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ will be;

$$\hat{y} = 2.9969594 + 0.126666x$$

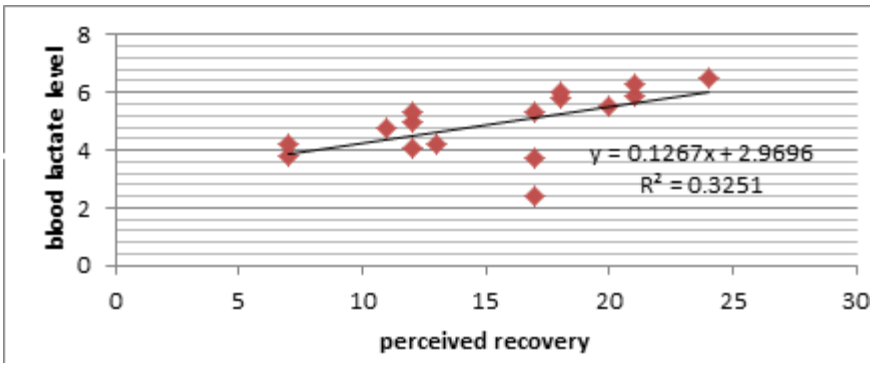

Figure 5: Scatterplot for data in Table 1

Comparing observed and predicted values

Table 3: Predicted values

y_i	Predicted $\hat{y} = 2.9969594 + 0.126666x$	Residual(error) $(y - \hat{y})$	Squared error $(y - \hat{y})^2$
3.8	3.856256	-0.05626	0.003165
4.2	3.856256	0.343744	0.11816
4.8	4.36292	0.43708	0.191039
4.1	4.489586	-0.38959	0.151777
5	4.489586	0.510414	0.260522
5.3	4.489586	0.810414	0.656771
4.2	4.616252	-0.41625	0.173266
2.4	5.122916	-2.72292	7.414272
3.7	5.122916	-1.42292	2.02469
5.3	5.122916	0.177084	0.031359
5.8	5.249582	0.550418	0.30296
6	5.249582	0.750418	0.563127
5.9	5.62958	0.27042	0.073127

6.3	5.62958	0.67042	0.449463
5.5	5.502914	-0.00291	8.49E-06
6.5	6.009578	0.490422	0.240514

The sum of the errors$(SE) = \sum(y - \hat{y}) = 0$ and $SSE = \sum(y - \hat{y})^2 = 12.65422$.

Table 4: Error of prediction

y_i	$(y_i - \bar{y}), \bar{y} = 4.925$	$(y_i - \bar{y})^2$
3.8	-1.125	1.265625
4.2	-0.725	0.525625
4.8	-0.125	0.015625
4.1	-0.825	0.680625
5	0.075	0.005625
5.3	0.375	0.140625
4.2	-0.725	0.525625
2.4	-2.525	6.375625
3.7	-1.225	1.500625
5.3	0.375	0.140625
5.8	0.875	0.765625
6	1.075	1.155625
5.9	0.975	0.950625
6.3	1.375	1.890625
5.5	0.575	0.330625
6.5	1.575	2.480625
		$SS_{yy} = \sum(y_i - \bar{y})^2 = 18.75$

From this we note that $SS_{yy}(18.75) > SSE(12.65422)$,so we can reason that x contributes data for the expectation of y

Estimating σ^2 (the variance)

$s^2 = \frac{SSE}{n-2} = \frac{12.65422}{16-2} = 0.903872$, therefore $s = 0.95$ is the estimated standard deviation.

Testing $H_0: \hat{\beta}_1 = 0$ vs $H_1: \hat{\beta}_1 \neq 0$

$\hat{\beta}_1 = 0.126666$ s = 0.95 and $SS_{xx} = 379.9375$ therefore

$$t = \frac{\hat{\beta}_1}{s/\sqrt{SS_{xx}}} = \frac{0.1266666}{0.95/\sqrt{379.9375}} = 2.59891772$$

$\alpha = 0.05, n = 16,\ v = (n-2) = 14.$ For a two tailed test, we reject $H_0$when $|t| > t_{\alpha/2,v}\ t_{0.025,14} = 2.145$

From the calculation above,we can see that $|t| > t_{\alpha/2,v}\ i.e, 2.599 > 2.145$, so the null hypothesis is rejested that is the slope $\hat{\beta}_1$ is not equal zero.

Coefficient of Determination r^2

$$r^2 = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{12.65422}{18.75} = 0.325108,\ \ therefore\, r = 0.57018$$

The values of r^2 shows about 57% of the sample variance in blood lactect concentration level can be explained by using the boxers perceived recovery x to predict blood lactect level y with the least square line $\hat{y} = 2.9969594 + 0.126666x$

At this point, lets use the the gotten regression model to predict the blood lactct level when perceived recovery (x) is 5, 15 ad 28

when x = 5

$\hat{y} = 2.9969564 + 0.126666(5) = 3.6$

when x = 15

$\hat{y} = 2.9969564 + 0.126666(15) = 4.9$

and when x = 28

$\hat{y} = 2.9969564 + 0.126666(28) = 6.5$

A prediction interval wher x = 5 with 95% prediction interval will be;

$$\hat{y} \pm (t_{\alpha/2}) s \sqrt{1 + \frac{1}{n} + \frac{(x_a - \bar{x})^2}{SS_{xx}}}$$

where $x_p = 5, \bar{x} = 15.375, n = 16, s = 0.95, SS_{xx} = 379.9375, t_{\alpha/2}$

$= 2.145$ with $v = n - 2$

$3.6 \pm (2.145)0.95\sqrt{1 + \frac{1}{6} + \frac{(5-15.375)^2}{379.9375}} = 3.6 \pm 2.453757 = (1.15, 6.05)$ which means the real value of y when the perceived recovery rate (x) is 4 should fall between 1.15 and 6.05.

Lets discuss a new concept the 'outliers'. These are simply points which appear unusual and far from other data points and they can be easily be seen or detected in a scatter plot.

Mathematicaly, data points greater than $Q_3 + 1.5(\text{IQR})$ and less than $Q_1 - 1.5(\text{IQR})$ where Q_1, Q_3 and IQR are the lower quantile, upper quantile and the inter quartile range rescptively are regarded as outliers.

So checking for outliers on the x-axis for our examples above is as follows , our x-values arranged in ascending order will be

7 7 11 12 12 12 13 17 17 17 18 18 20 21 21 24

Then median of the entire data $Q_2 = \frac{17+17}{2} = 17$

the median of the lower half of the data $Q_1 = \frac{12 + 12}{2} = 12$

and themedian of the upper half of the data $Q_3 = \frac{18+20}{2} = 19$

therefore $\text{IQR} = Q_3 - Q_1 = 7$

So our outliers will be data points less than;

$$12 - 1.5(7) = 1.5$$

And data points greater than

$$19 + 1.5(7) = 29.5$$

It seems there is no oulier in our data as all data fall with the range of (1.5, 29.5) and this also tells us that our regresioon model becomes useless to predict future values of y as x gets smaller than 1.5 and greater than 29.5.

Chapter 3

EXPERIMENTAL ANALYSIS

As stated earlier, manual computations of the various regression coefficients is a difficult task when dealing with data of big observations. However, with the invention of statistical software one of which is SPSS has made it a lot easier for scientist to statisticians to carry out regression analysis with ease. Here we will be using SPSS in carrying out regression analysis on the number of visitors visiting selected departments of the health center.

The data below shows the number of visitors visiting the school health center and the purpose for their visit, for the year 2014 and 2015.

Table 5: Number of visitors that visited the school health center and the department they visited between January 2014 to October 2015

MONTH/YEAR	INTERNAL MEDICINE	DERMATOLOGY	PSYCHIATRY	DENTAL	OPHTHALMOLOGY	GYNOCOLOGY	EAR-NOSE-THROAT	TOTAL
January,2014	508	160	48	225	268	138	791	2812
February, 2014	438	105	37	172	191	84	297	1535
March,2014	641	266	64	240	388	138	916	3451
April,2014	706	204	49	245	437	182	1095	4001
May,2014	612	235	65	300	315	161	916	3300
June,2014	359	134	51	288	186	153	488	1911
July,2014	206	69	15	127	88	53	197	844
August,2014	257	75	20	159	77	21	226	904
September,2014	187	100	31	156	108	53	259	1005
October,2014	445	161	48	234	242	117	616	2326
November,2014	671	237	46	261	343	182	957	3499
December,2014	855	225	54	290	359	202	1087	4003
January,2015	468	143	54	210	199	162	81	1244
February, 2015	370	85	28	164	143	93	85	947
March,2015	562	239	39	272	309	188	1102	3611
April,2015	712	234	74	244	431	185	1472	4777
May,2015	637	215	43	236	331	208	964	3478
June,2015	5	220	48	248	277	204	455	1721
July,2015	1	56	22	163	138	85	380	1144
August,2015	326	77	33	101	93	42	406	1399
September,2015	5	220	48	248	277	204	1135	3081
October,2015	518	229	48	258	376	118	964	3364

The above information was gotten from the school health center secretary's office where all visitor must register before they could see a doctor .this was carried out for orderliness and to render the best possible health service for all visitors without stress.

Here our duty is to carry out a regression analysis on some key departments of the health center to know the number of visitor they attend to in a given period and to see which one these department need more man power.

3.1 A Regression Analysis Between the Total Number of Visitors and those Visiting Ear-Nose and Throat

Figure 7: Descriptive Statistics

	Mean	Std. Deviation	N
ENT	676.7727	405.00405	22
Total	2470.7727	1235.82777	22

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.973[a]	.946	.944	96.18291

ANOVA[a]

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	3259570.812	1	3259570.812	352.342	.000[b]
	Residual	185023.052	20	9251.153		
	Total	3444593.864	21			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	-110.899	46.705		-2.374	.028
	Total	.319	.017	.973	18.771	.000

Therefore the regression model is ;

$$y = -110.899 + 0.319x$$

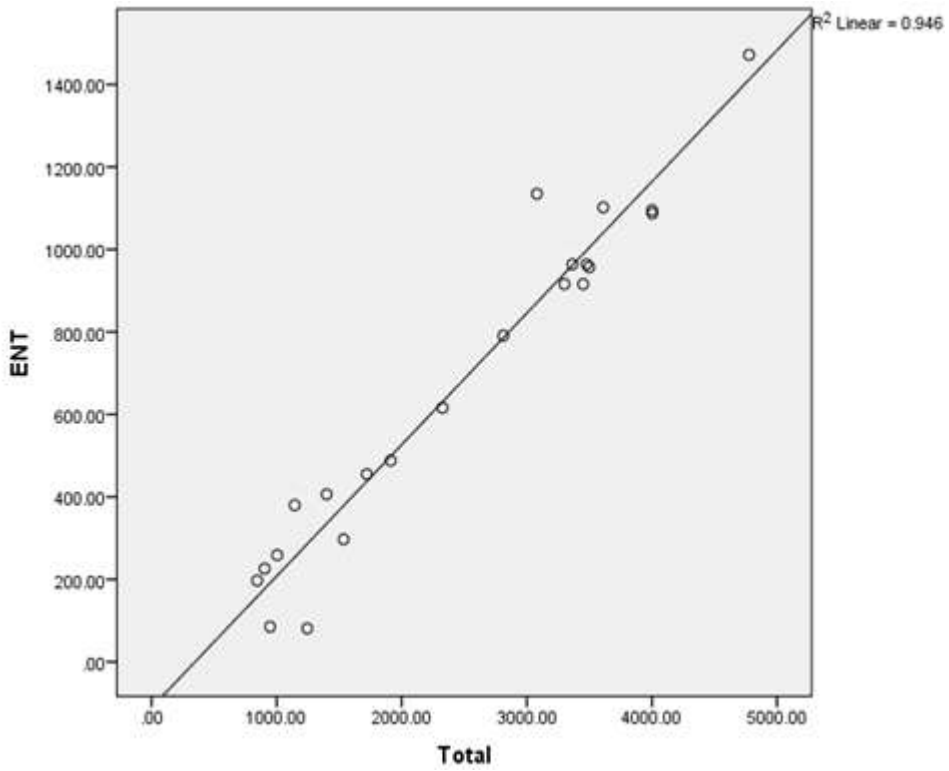

Figure 8: Scatterplot for Ear-Nose-Throat against the total number of visitors

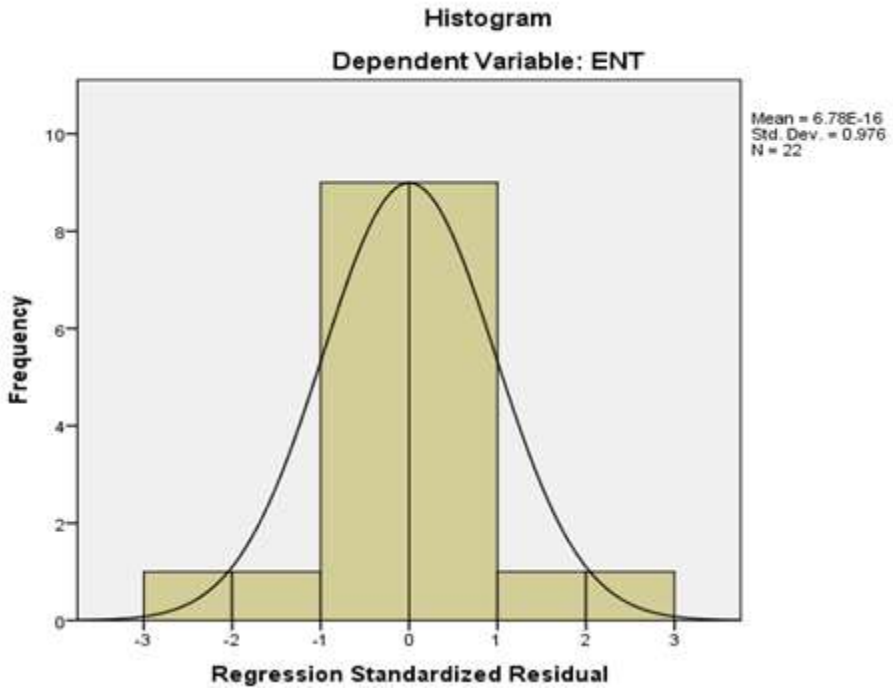

Figure 9:Normal Distribution of Ear-Nose-Throat against Total number of Visitors

3.2Analysis of Total Number of Visitor and those Visiting the Dermatological Department

Here we shall be considering the number of visitor visiting the the dermatological department of the health center. Using the data in table 3.1, we have the following outcomes;

Figure 10: Descriptive Statistics

	Mean	Std. Deviation	N
Dermatology	167.6818	69.27611	22
Total Visitor	2470.7727	1235.82777	22

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.878[a]	.771	.760	33.96656

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	77708.227	1	77708.227	67.354	.000[b]
	Residual	23074.546	20	1153.727		
	Total	100782.773	21			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	46.064	16.494		2.793	.011
	Total	.049	.006	.878	8.207	.000

Therefore the regression model is;

$$y = 46.064 + 0.049x$$

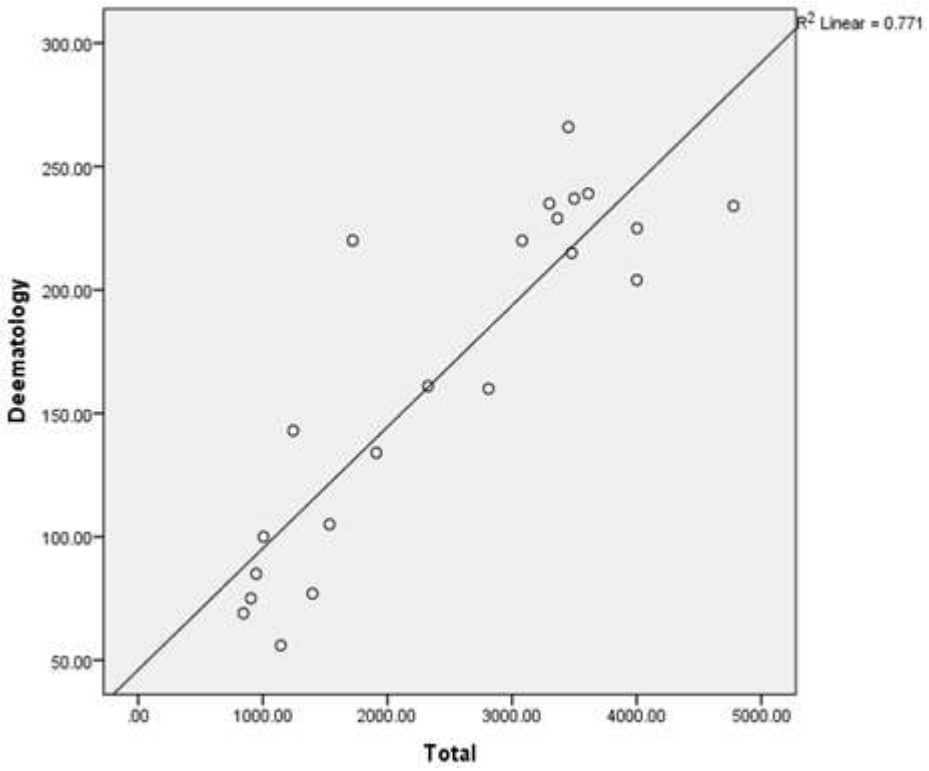

Figure 11: Scatterplot for Dermatology and Total number of Visitors

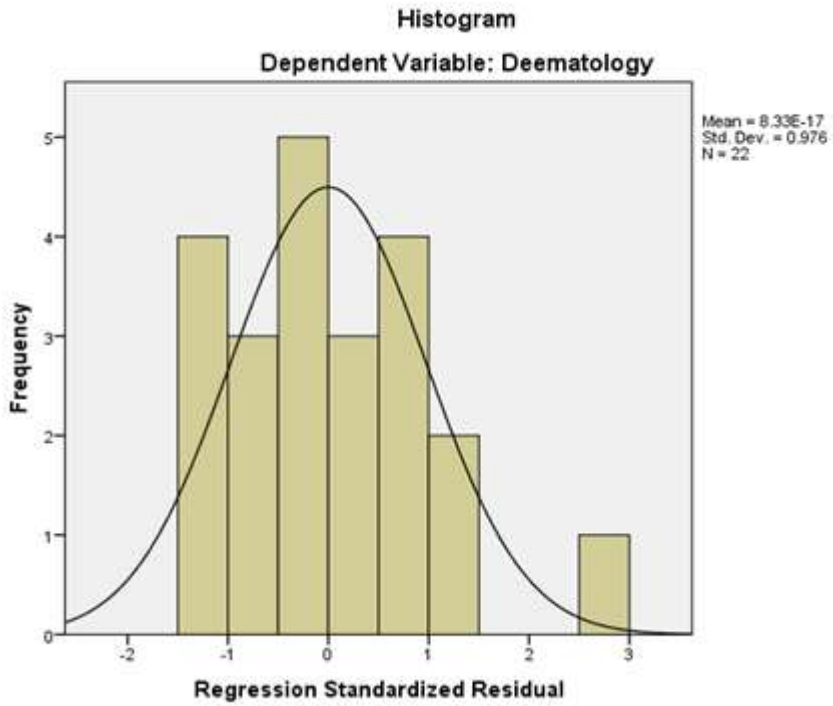

Figure 12: Normal Distribution of Dermatology against Total number of Visitors

3.3 Analysis on Visitors Visiting Ophthalmological Department

Figure 13: Descriptive Statistics

	Mean	Std. Deviation	N
Ophthalmological	253.4545	114.26524	22
Total	2470.7727	1235.82777	22

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.943[a]	.889	.884	38.94972

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	243845.838	1	243845.838	160.734	.000[b]
	Residual	30341.617	20	1517.081		
	Total	274187.455	21			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	38.016	18.913		2.010	.058
	TotalVisitor	.087	.007	.943	12.678	.000

Therefore the regression is;

$$y = 38.016 + 0.087x$$

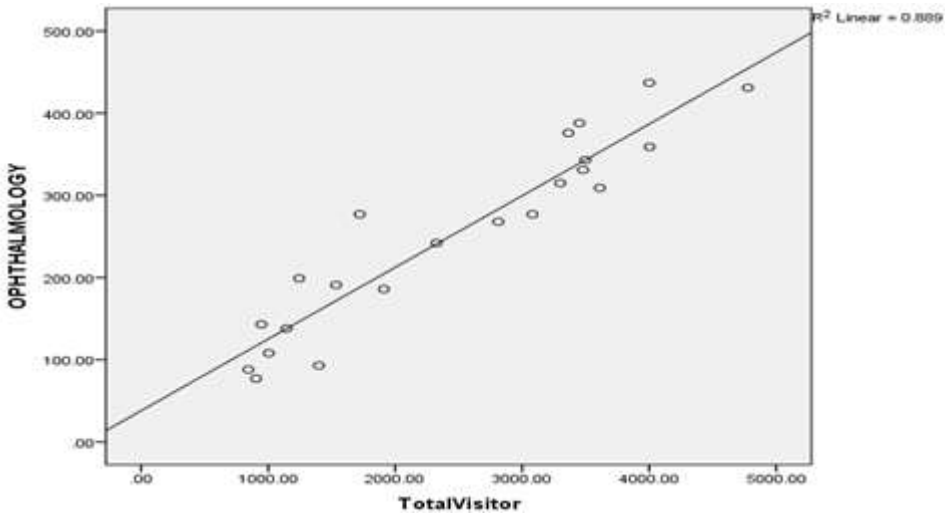

Figure 14: Scatterplot of Ophthalmological againstTotal number of Visitors

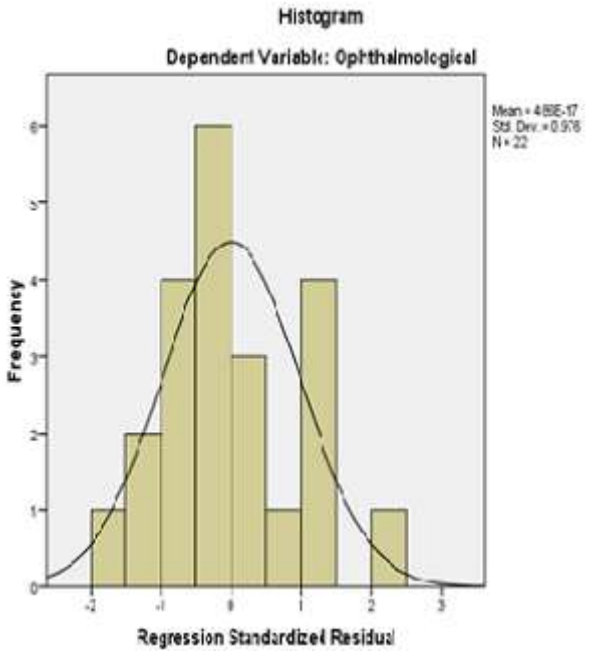

Figure 15: Normal Distribution of Ophthalmology against Total number of Visitors

3.4 Analysis on Visitors Visiting the Health Center Monthly

Figure 16: Descriptive Statistics

	Mean	Std. Deviation	N
Monthly	11.5000	6.49359	22
Total	2470.7727	1235.82777	22

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.033[a]	.001	-.049	6.65024

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.987	1	.987	.022	.883[b]
	Residual	884.513	20	44.226		
	Total	885.500	21			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	11.067	3.229		3.427	.003
	Total	.000	.001	.033	.149	.883

Therefore the regression model is ;

$$y = 11.067$$

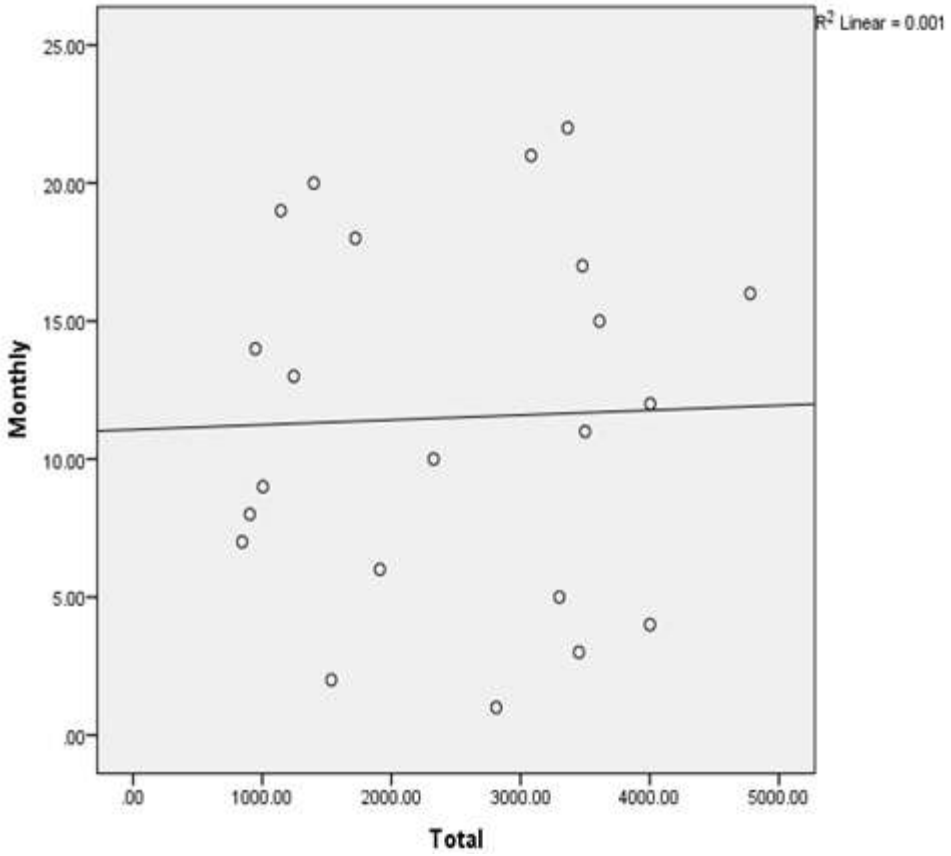

Figure 17: Scatterplot of Monthly visitors against Total number of Visitors

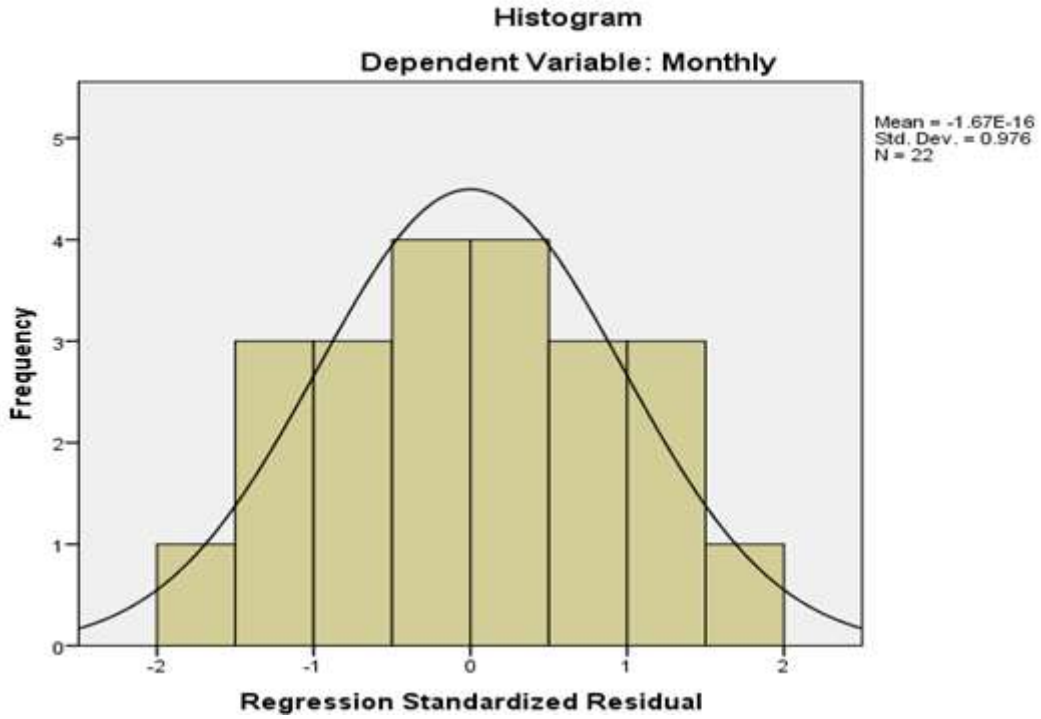

Figure 18: Normal Distribution of Monthly Visitors against Total number of Visitors

Chapter 4

RESULT AND DISCUSSION

The mean and the standard deviation of visitor visiting the Ear-Nose-Throat department are 676.7727 and 405.00405 respectively while the regression equation associated with this department is y=-110.899+0.319x having a linear relationship defining on R=0.973 which means that about 97% of the total visitors visiting the health center visit the Ear-Nose-Throat department, this shows a very high correlation with a normal standard error of 96.18291.

Assuming the health center had a total number of 2500 visitors in a month then about y=-110.899+0319(2500)=686.601 which is approximately 687 visitors will be visiting this department.

The analysis from section 3.3 shows that the dermatological department has a mean and standard deviation of visitors of 167.6818 and 69.27611 respectively with a regression mode given as y=46.064+0.049x having a liner relationship of R=0.878 which tells us that about 88% of the visitors visiting the health center visit this department and this is also a high correlation with a normal standard error of 33.96656.

Predicting the number of visitors who will visit the school dermatologist when a total number of 2500 visitors visit the health center will be y=46.064+0.049(2500)=168.564 which is about 169 visitors.

Also the mean and standard deviation of those visiting the Ophthalmological department are 253.4545 and 114.26524 respectively with a linear regression model of y=38.016+0.087x with a linear relationship of R=0.943 and R^2=0.889 which shows that about 94% of the total visitors a month visits this department with a standard error of 38.94972.

Predicting the number of those that will visit this department if the health centers has a total visitors of 2500 will be about y=38.016+0.087(2700)=255.5 and its approximately 256 visitors.

Lastly, the mean and standard deviation of the number of visitors visiting the health center monthly are 11.5 and 6.49359 with a regression model of y=11.067 having a correlation of R=0.03n and R^2=0.001 and this shows that there are not linearly correlated but have a non-linear correlation as can be seen in figure 13 and figure 14.

4.1 Discussion

From the above results, it shows clearly that visitors visiting the the school health center often visit the department of Ear-Nose-Throat, the dermatological department and the Ophthalmological department with Ear-Nose-throat taking the lead next by Ophthalmological department and finally Dermatological department .

Based on this, we urge the school management/health management to as a matter of urgency look in on how to improve the efficiency of staffs in the said departments particularly the department of Ear-Nose-Throat which currently have just one medical personal. By employing more professionals and als carry out a research on the cause of the high number of these cases so as to to come up with a preventive measures for a healthy and safe EMU.

REFERENCES

[1] Wilhox, Waiter (1938) The Founder of Statistics. *Review of the international Statistical Institute* 5(4):321-328 JSTOR 1400906.

[2] Galton, F(1877). Typical Laws of Heredity Nature *15:492-553 doi:10.1038/015492a0*

[3] A.M. Legendre (1805) Nouvelles methods pour la determination des orbites des cometes. Firmin Didot, Paris,. *Sur la Methode des Morindres quarres appears as an appendix*

[4] C.F. Gauss (1809). Theoria Motus Corporum Coalestium in Sectionibus Conicis Solem Ambientum.

[5] William Mendenhall and Terry Sincich. A second course statistics Regression Analysis.