MULTIPLE LINEAR REGRESSION ANALYSIS OF MATHEMATICS, ENGLISH, K.C.P.E AND KCSE SCORES IN NDHIWA SUB – COUNTY, KENYA

By

DIMBA KENNEDY OKUKU

A project report

Submitted in partial fulfillment

of the requirements for the degree of

MSC in Applied Statistics.

School of Mathematics, Statistics and Actuarial Science

MASENO UNIVERSITY

Abstract

The Kenyan ministry of education entirely uses the KCPE aggregate marks to selectively admit students to form one. This criterion may be conceptualized in terms of errors of false acceptance and false rejection. Because prediction is imperfect, some students admitted in form one will fail in their KCSE (false acceptance) and others who are rejected would succeed if given the opportunity (false rejection).

This research used the method of least squares to come up with a statistical model that can be used to estimate the KCSE scores based on the KCPE marks, Mathematics scores and the English marks. The model used assumed a linear relationship between the variables; the error term was also assumed to have constant variance, expected value of zero and normally distributed.

This research will help the ministry of education to evaluate the efficiency of their criterion of form one selection and thus take appropriate measures based on the results of the research. It will also ensure that deliberate efforts are made to prepare adequate facilities and capacities needed for transition from secondary schools to tertiary institutions.

The research methodology used involved both random and stratified sampling of the secondary data obtained from the KCPE results analysis of the target county secondary schools and sub – county schools. The population was stratified into county and sub – county from which a sample was drawn randomly. In Multiple linear regression analysis spss software was used to come up with the parameter estimates and to do other statistical analysis.

The analysis of the data revealed that the KCPE aggregate score, Mathematics scores and the English score are statistically significant at 0.05 significance level. We therefore concluded that indeed three predictor variables are important in the prediction of the KCSE results. Hence the sub – county should consider incorporating the three predictor variables in their form one selection.

There were also certain issues that this research was unable to address due to its scope such school administration, teacher – student ratio, school infrastructures and resources which inevitably causes variation in the KCSE performance of different students. These are areas where more research should be considered. Also English should be given a lot of emphasis since it is the medium of instruction being used in both primary and secondary schools in the Kenyan schools.

CHAPTER ONE: INTRODUCTION

1.1 Background information of the study

Education is one of the key pillars of vision 2030. Therefore provision of quality and relevant education cannot be overestimated. As it will ensure production of human capital with the skills required for the realization of the vision 2030.

The Kenyan ministry of education uses the aggregate marks scored in KCPE to allot students to the limited chances available in the national, county and sub - county schools. The ministry hardly considers any other factor other than the KCPE marks locking out several students from getting admissions into their dream schools. Since admissions to secondary schools, and particularly county schools remains limited and competitive, the question of admissions policy is a perennial one. Conflicting government, public and other education stakeholders interests exist in attempting to balance primary – secondary schools transition rate against the performance expectations required of those pursuing secondary education.

In Kenya, the primary determinant of admission to secondary schools is the aggregate KCPE marks. This admission criterion may be conceptualized in terms of errors of false acceptance and false rejection. Because prediction is imperfect, some students admitted will fail (false acceptances) and others who are rejected would succeed if given the opportunity (false rejection).

The ministry of education more often than not varies the minimum entry behaviour into the tertiary institution due to the constraints of the facilities and human capital in the tertiary institutions .It is against this background that we explored a statistical model that was used to estimate the scores of students in KCSE based on the aggregate KCPE marks, Mathematics and English scores. A multiple regression analysis with three variables was used. Multiple regression is a flexible method of data analysis appropriate whenever dependable variable is to be examined in relationship to the independent variable. Regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is the average value of the dependent variable when the independent variable is fixed. We also wish to investigate the statistical significance of each of the independent variables.

1.2 Statement of the problem

During form one selection the ministry of education normally uses the aggregate KCPE marks to post the candidates to the limited and competitive chances in the form one classes. There has been an outcry for the review of this criterion to adopt a selection process that focuses on broad access to secondary education as opposed to high selectivity as this would ensure a largely literate society.

Accurate predictions can be realized through the use of statistical model that considers the major components of the aggregate KCPE marks.

This research attempted to come up with a statistical model that incorporates Mathematics, English and the aggregate KCPE marks in the form one selection. It has also further determined the contributions of each of these three components to the KCSE scores.

1.3 Objectives of the study

The main objectives of this study are:

(i) To establish the relationship between KCPE aggregate scores, Mathematics score, English score and the KCSE scores.

(ii) To determine whether KCPE aggregate scores, Mathematics scores and the English marks taken together as a group are useful in predicting the KCSE scores.

(iii) To determine which of the three predictor variables KCPE aggregate score, Mathematics score, and English score is more useful in predicting the KCSE scores.

1.4 Significance of the study

(i) The model may be used by the ministry of education and other education stakeholders to revamp facilities in the primary schools to ensure better performance in secondary schools.

(ii) The research may help the ministry of education to evaluate its criteria of form one selection largely based on the K.C.P.E marks and consider other options.

(iii) The results from the study will also help the government prepare adequate facilities for those to join tertiary institutions as they will be able to forecast the grades to be scored by students.

CHAPTER TWO

LITERATURE REVIEW

Quite a number of scholars have come up with regression models:

Legendre in 1805 and Gauss 1809 both applied the method of least squares, to determine from astronomical observations the orbits of the bodies about the sun. Gauss published a further development of the theory of least squares in 1821.

The term regression was first coined by Francis Galton in the 19th century to describe a biological phenomenon. The phenomenon was that the heights of the descendants of the tall ancestors tend to regress down towards a normal average, also known as regression towards the mean

For Galton, regression hard only biological meaning but this work was later extended by Udry Yule and Karl Pearson to a more general context. In their work of the joint distribution of the response and the explanatory variables are assumed to be Gaussian.

This assumption was weakened by R.A Fisher in work of 1922 and 1925.Fisher assumed that the conditional distribution of the response variable is Gaussian but the joint distributions need to be in this respect. Fisher's assumption is closer to Gauss formulation of 1921.

Regression continues to be an area of active research. In recent decades new methods have been developed for robust regression involving correlated responses such as time series and growth curves, regression in which the predictor or response variable are curves, images or graphs or other complex data objects, regression methods accommodating various types of data.

Hallack and Poison (2007) [16]states that the main function of public examinations is to serve as instruments for making objective and neutral judgment. According to a study by World Bank 2007 learning assessment are crucial for measuring education quality and relevant diagnosing system, weaknesses and motivating policy reforms.

A study by Fuller (1987), Gwewwe and Kremer (2006) [10]shows a positive relationship between students entry marks and performance. Also a study by wanjohi and Yara (2011) [23]category of school predicts performance of students in KCSE.

Andrian (2008)[1] asserts that many problems that bedevil students in high schools have their roots from the primary level of education. He argues that English as a subject plays a key role in students' performance as besides being an examinable subject it is the language of instruction in secondary schools.

According to Lewin (2008)[19] Primary examination is largely content rather than skill based and reward recall rather than the higher cognitive capabilities characterized by the secondary

examination. This he says could be the reason for the differential performance between KCPE and KCSE examinations.

Public exit examinations can provide valuable information which can hold both schools and students accountable (Hunshek 2003)[17]. Students in countries with public exit examinations systems tend to be systematically outperform countries without such systems

Weak monitoring and assessment systems remains major obstacle for improved learning outcomes at the secondary school level (Bregman and stallmaster, 2002). Systematic and internationally comparable assessment of learning in secondary education at classroom is not widespread and considerable reliance has been placed on public examination to ensure the common curricula are covered.

The literature review indicates that little or no studies have been done in Kenya particularly Ndhiwa sub – county to use mathematics, English and KCPE scores to predict the KCPE scores to predict the KCSE performance, this research seeks to fill this gap.

CHAPTER THREE: Basic mathematical concepts

3.1 Introduction

This chapter reviews some of the mathematical concepts and definitions required as a background in building this project.

The general purpose of multiple regression analysis is to quantify the relationship between several predictor variables and dependent variable. It can also be used to estimate the effect of predictor variable on the dependent variable.

3.2 Derivation OLS estimates

Consider a three variable multiple linear regression model. Then the regression population regression equation (PRE) is given as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \mu_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where μ_i is iid random error term. The OLS sample regression equation (OLS – SRE) for equation (1) is

$$Y_i = \beta`_0 + \beta`_1 X_{1i} + \beta`_2 X_{2i} + \mu`_i$$
$$= Y`_i + \mu`_i$$

Where:

$i = 1 \ldots\ldots\ldots\ldots\ldots\ldots N$

β_j= OLS estimators of the corresponding population regression co – efficient β_j (j = 0,1 2).

The ordinary least square residuals is given by:

$$\mu`_i = Y_i - Y`_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

$$= Y_i - \beta`_0 - \beta`_1 X_{1i} - \beta`_2 X_{2i}$$

(i = 1....N)

The predicted values of Y_i is given by

$$Y`_i = \beta`_0 + \beta`_1 X_{1i} + \beta`_2 X_{2i} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

(i = 1.....N)

The ordinary least square sample regression function is given by

$$f (X_{1i}, X_{2i}) = \beta`_0 + \beta`_1 X_{1i} + \beta`_2 X_{2i} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

Derivation of the ordinary least squares co – efficient estimates.

The sample regression equation is given by

$$Y_i = \beta`_0 + \beta`_1 X_{1i} + \beta`_2 X_{2i} + \mu`_i$$

$= Y_i - \beta`_0 - \beta`_1 X_{1i} - \beta`_2 X_{2i}$ ..... (5)

Squaring both sides and summing over N

$\sum (\mu`_i)^2 = \sum (Y_i - \beta`_0 - \beta`_1 X_{1i} - \beta`_2 X_{2i})^2$

$S = \sum (Y_i - \beta`_0 - \beta`_1 X_{1i} - \beta`_2 X_{2i})^2$ ..... (6)

Where $S = \sum (\mu`_i)^2$

So we seek to find the values of $\beta`_0$, $\beta`_1$, $\beta`_2$ that minimizes the quantity S. To achieve this we seek the partial derivatives of S with respect to the critical values $\beta`_0$, $\beta`_1$, $\beta`_2$ and set each of the derivatives to zero

$\partial S/\partial \beta`_0 = -2 \sum \mu`_i$

$= \sum (Y_i - \beta`_0 - \beta`_1 X_{1i} - \beta`_2 X_{2i}) = 0$ ..... (7)

$\partial S/\partial \beta`_1 = -2 \sum X_{1i} \mu`_i = 0$

$= \sum X_{1i} (Y_i - \beta`_0 - \beta`_1 X_{1i} - \beta`_2 X_{2i}) = 0$ ..... (8)

$\partial S/\partial \beta`_2 = -2 \sum X_{2i} \mu`_i = 0$

$= \sum X_{2i} (Y_i - \beta`_0 - \beta`_1 X_{1i} - \beta`_2 X_{2i}) = 0$ ..... (9)

N/B

In the above partial differentiation (eqn 7 – 9) we have moved the summation operator ($\sum$) out of front since derivative of a sum is equal to the sum of derivatives.

Hence the normal equations are gotten by taking summations and rearranging terms

- $\sum (Y_i - \beta`_0 - \beta`_1 X_{1i} - \beta`_2 X_{2i}) = 0$

  $\sum Y_i - N\beta`_0 - \beta`_1 \sum X_{1i} - \beta`_2 \sum X_{2i} = 0$

  $-N\beta`_0 - \beta`_1 \sum X_{1i} - \beta`_2 \sum X_{2i} = - \sum Y_i$

  $N\beta`_0 + \beta`_1 \sum X_{1i} + \beta`_2 \sum X_{2i} = \sum Y_i$ ..... (10)

- $\sum X_{1i} (Y_i - \beta`_0 - \beta`_1 X_{1i} - \beta`_2 X_{2i}) = 0$

  $\sum X_{1i} Y_i - \beta`_0 \sum X_{1i} - \beta`_1 \sum X_{1i}^2 - \beta`_2 \sum X_{1i} X_{2i} = 0$

$$-\beta`_0\sum X_{1i} - \beta`_1\sum X_{1i}^2 - \beta`_2\sum X_{1i}X_{2i} = -\sum X_{1i}Y_i$$

$$\beta`_0\sum X_{1i} + \beta`_1\sum X_{1i}^2 + \beta`_2\sum X_{1i}X_{2i} = \sum X_{1i}Y_i \quad \text{.....(11)}$$

- $\sum X_{2i}(Y_i - \beta`_0 - \beta`_1X_{1i} - \beta`_2X_{2i})=0$

$$\sum X_{2i}Y_i - \beta`_0\sum X_{2i} - \beta`_1\sum X_{1i}X_{2i} - \beta`_2\sum X_{2i}^2 = 0$$

$$-\beta`_0\sum X_{2i} - \beta`_1\sum X_{1i}X_{2i} - \beta`_2\sum X_{2i}^2 = -\sum X_{2i}Y_i$$

$$\beta`_0\sum X_{2i} + \beta`_1\sum X_{1i}X_{2i} + \beta`_2\sum X_{2i}^2 = -\sum X_{2i}Y_i \quad \text{.....(12)}$$

Assemble the three OLS normal equations (10 –12)

$$N\beta`_0 + \beta`_1\sum X_{1i} + \beta`_2\sum X_{2i} = \sum Y_i$$

$$\beta`_0\sum X_{1i} + \beta`_1\sum X_{1i}^2 + \beta`_2\sum X_{1i}X_{2i} = \sum X_{1i}Y_i$$

$$\beta`_0\sum X_{2i} + \beta`_1\sum X_{1i}X_{2i} + \beta`_2\sum X_{2i}^2 = -\sum X_{2i}Y_i$$

We then proceed and find the solution of the normal equations N1 ...N3 which yields explicit expressions for $\beta`_0$, $\beta`_1$, $\beta`_2$. These expressions are the OLS estimator's $\beta`_0$, $\beta`_1$, $\beta`_2$ of the partial regression co – efficient β_0, β_1, β_2 respectively.

We define the deviations from the means of Y_i, X_{1i}, and X_{2i} as:

$y_i = Y_i - \bar{Y}$

$x_{1i} = X_{1i} - \bar{X_1}$

$x_{2i} = X_{2i} - \bar{X_2}$

Where:

$\bar{Y} = \sum Y_i/N$ is the sample mean of Y_i.

$\bar{X}_1 = \sum X_{1i}/N$ is the sample mean of the X_{1i} values.

$\bar{X}_2 = \sum X_{2i}/N$ is the sample mean of the X_{2i} values

The OLS co – efficient estimator $\beta`_1$ and $\beta`_2$ deviation from means form are:

$$\beta`_1 = \{ (\sum_i x_{2i}^2)(\sum_i x_{1i}y_i) - (\sum_i x_{1i}x_{2i})(\sum_i x_{2i}y_i)\} /\{ (\sum_i x_{1i}^2) (\sum_i x_{2i}^2) - (\sum_i x_{1i}x_{2i})^2\} \quad \text{.....(13)}$$

$$\beta`_2 = \{ (\sum_i x_{1i}^2) (\sum_i x_{2i}y_i) - (\sum_i x_{1i}x_{2i}) (\sum_i x_{1i}y_i)\} / \{ (\sum_i x_{1i}^2) (\sum_i x_{2i}^2) - (\sum_i x_{1i}x_{2i})^2\} \text{.....(14)}$$

$$\beta`_0 = \bar{Y} - \beta`_1 X_1 - \beta`_2 X_2 \text{.....(15)}$$

3.3 The ordinary least square variance – covariance estimators

An unbiased estimator of the error variance σ^2.

For a general multiple linear regression model with k regression co – efficient, an unbiased estimator of the error variance σ^2 is the degree of freedom adjusted estimator

$\sigma^{`2} = \sum_i \mu^{`2}/N - K$

$= RSS/ N - K$

Where $K = k + 1$

= The total number of regression co – efficient in equn 1

RSS = Residual Sum of squares

For $K = 3$

$\sigma^{`2} = \sum_i \mu^{`2}/ N - 3$

$= RSS/N - 3$

Formulas for the variance and covariances of the slope co – efficient estimators $\beta`_1$ and $\beta`_2$ in the three variable multiple regression model.

$$Var(\beta`_1) = \sigma^2 \sum_i x_{2i}^2 / \{(\sum_i x_{1i}^2)(\sum_i x_{2i}^2) - (\sum_i x_{1i}x_{2i})^2\}$$

$$Var(\beta`_2) = \sigma^2 \sum_i x_{2i}^2 / \{(\sum_i x_{1i}^2)(\sum_i x_{2i}^2) - (\sum_i x_{1i}x_{2i})^2\}$$

$$Cov(\beta`_1, \beta`_2) = \sigma^2 \sum_i x_{1i} x_{2i} / \{(\sum_i x_{1i}^2)(\sum_i x_{2i}^2) - (\sum_i x_{1i}x_{2i})^2\}$$

The unbiased estimators of the variances of the slope of the slope of the co – efficient estimators $\beta`_1$ and $\beta`_2$ are obtained by substituting the unbiased estimator $\sigma^{`2}$ for the unknown error variance $\sigma^2$ in the formulae for Var $(\beta`_1)$ and Var $(\beta`_2)$

$$\text{Var}(\beta`_1) = \sigma`^2 \sum_i x_{2i}^2 / \{(\sum_i x_{1i}^2)(\sum_i x_{2i}^2) - (\sum_i x_{1i}x_{2i})^2\}$$

$$\text{Var}(\beta`_2) = \sigma`^2 \sum_i x_{2i}^2 / \{(\sum_i x_{1i}^2)(\sum_i x_{2i}^2) - (\sum_i x_{1i}x_{2i})^2\}$$

$$\text{Cov}(\beta`_1, \beta`_2) = \sigma`^2 \sum_i x_{1i}\, x_{2\,i} / \{(\sum_i x_{1i}^2)(\sum_i x_{2i}^2) - (\sum_i x_{1i}x_{2i})^2\}$$

3.4 Goodness of fit of multiple linear regression

Goodness of fit attempts to get how well a model fits a given set of data or how well it will predict future set of observations. The assumptions of a model may not perfectly hold however the assumptions may hold closely enough for the model to be useful in practice. The assumptions must hold closely enough to allow us make predictions or make inferences about the effects of one variable on another.

The following concepts can be used to measure the goodness of fit:

(i) Examining the residuals from the model.
(ii) Outlier detection
(iii) A global measure of variance` explained` R^2.
(iv) A global measure of variance explained that is adjusted for the number of parameter in a model, adjusted R^2.

Examining residuals from the model

This is the most informative methods to investigate a model fit. The histograms or scatter diagrams are examined to investigate the model fit. The histogram should have a normal shape and the scatter plots should show little departure from constant variance or linearity.

Outlier detection

An outlier is an extreme observation, different from other observations in the data set. They are identified from the residual plots. Upon detection of outliers the data set should be verified for error, or the outlying observation may be deleted.

R^2: A global measure of variance explained

We may also use the co – efficient of multiple determination R^2 as a global statistic to assess the fit of a model. The R^2 is routinely given in the R software outputs. It reveals the worth of the independent variable. It is computationally given as

$$R^2 = SS_R/SS_T$$

$$=1 - SS_E/SS_T$$

Where:

SS_R = Sum of squares due to regression.

SS_T = Sum of squares due to total.

SS_E = Sum of squares due to error.

By definition of least square regression $SS_E \leq SS_T$, so $0 \leq R^2 \leq 1$. If $SS_E = SS_T$ the $R^2 = 0$ and the model is not useful. If $SS_E = 0$, then $R^2 = 1$ and the model fits all the points perfectly. Almost all points will be between these extremes.

The square root of R^2 gives correlation co – efficient between dependent and the explanatory variables, except possibly for the sign which is lost in taking the square. As a number of regression co – efficients are added R^2 never goes down even if the additional explanatory variable is not useful i.e. there is no adjustment in the number of parameters in the model.

Adjusted R^2

Many regression users prefer to use the adjusted R^2 statistics

$$R^2_{Adj} = 1 - \{[(SS_E/(n- k)]/SS_T(n - 1)\}$$

$$=1 - \{[(n - 1)\, SS_E]/\, [SS_T\,(n - k)]\}$$

Where:

n = sample size

k = Number of variables

The adjusted R^2 statistic essentially penalizes the analyst for adding terms to the model. It is an easy way to guard against over fitting, which is including regressors' that are not really useful. Consequently it is useful in computing and evaluating competing regression models.

3.5 Test for significance of regression

The test for significance of regression is a test to determine whether a relationship exists between the response variable y and a subset of regressor variables

$x_1, x_2, \ldots\ldots\ldots x_k$. The appropriate hypothesis:

$H_0: \beta_1 = \beta_2 = \ldots\ldots\ldots\ldots\ldots\ldots = \beta_k = 0$

$H_1: \beta_j \neq 0$ for at least one j

Rejection of $H_0: \beta_1 = \beta_2 = \ldots\ldots\ldots\ldots\ldots\ldots = \beta_k = 0$ implies that at least one of the regressor variables $x_1, x_2, \ldots\ldots\ldots x_k$ contributes significantly to the model.

The total sum of squares SST is partitioned into a sum of squares of squares due to regression and sum of squares due to error say

$SS_T = SS_R + SS_E$

If $H_0: \beta_1 = \beta_2 = \ldots\ldots\ldots\ldots\ldots\ldots = \beta_k = 0$ is true then

We find a computing formula for SSE as follows

$SSE = \sum_{i=1}(yi - \tilde{y})^2$

$= \sum_{i=1} e_i^2$

$= e`e$ ..... *

Substituting $e = y - \tilde{y}$

$= y - x\beta`$

Into the (*) we obtain

$SSE = Y`Y - \beta`X`Y$ .....**

Analysis of variance for testing significance of regression in multiple regression

Source of variation	Sum of squares	Degree of freedom	Mean square	f_o
Regression	SS_R	K	MS_R	MS_R/MS_E
Error or residual	SS_E	n-k-1	MS_E	
Total	SS_T	n-1		

Now since

$SST = \sum_{i=1} y_i^2 - (\sum_{i=1} y_i)^2 / n$

$= yy` - (\sum_{i=1} y_i)^2 / n$

We write equation ** as

$SSE = y`y - (\sum_{i=1} y_i)^2 / n - [\beta`X`Y - (\sum_{i=1} y_i)^2 / n]$

$= SS_T - SS_R$

Therefore the regression sum of squares is

$SS_R = \beta`X`Y - (\sum_{i=1} y_i)^2 / n$

3.6 Model specification

In studying the relationship between KCSE scores and aggregate KCPE marks, Mathematics and English marks, the study used a multiple linear regression analysis which assumed that the response variable y_i (KCSE scores) is related to the explanatory variable x_1 (KCPE marks) x_2 (Mathematics marks) and x_3(English marks) by:

$Y` = \beta`_o + \beta`_1 x_{1i} + \beta`_2 x_{2i} + \beta`_3 x_{3i} + \varepsilon`_i$

Where:

i= 1, 2.....n

ε_i= Unknown error component superimposed on the true linear relation. These may include factors such as learning environment, school leadership etc.

$\beta`_o$ = Regression constant

$B`_1, \beta`_2, \beta`_3$ are the regression co- efficient for the students' scores in aggregate KCPE, Mathematics and English respectively.

x_{1i}, x_{2i}, x_{3i} are the i^{th} KCPE aggregate, Mathematics and English scores respectively.

The above was based on the following assumptions

(i) The sample is representative of the population for inference prediction.

(ii) The residuals are normally distributed with a mean of zero and a constant variance. A histogram of the residuals was drawn to determine this.

(iii) Linearity

The independent variable is assumed to be linearly independent i.e it is not possible to express it is not possible to express any independent variable as a linear combination of others. If this assumption is not met then prediction may be systematically overestimate the actual values for one range of values on the predictor variable and overestimate them for another.

(iv) Homoscedacity

It is assumed that the variance of o the errors of prediction are the same for all the prediction values.

CHAPTER FOUR: RESEARCH METHODOLOGY

4.1 Introduction

This chapter provides information on the area of study, population, sampling techniques, research design data collection methods and data analysis.

4.2 Research Design

This was across sectional study aimed at analyzing the relationship between KCSE scores and the KCPE aggregate marks and mathematics and English scores. There are two categories of schools in Ndhiwa the county schools and the sub – county schools. Due to this heterogeneity randomized complete bock design was the most appropriate and the sample size was drawn from these sampling units (strata).Stratified random sampling was used as it increases the efficiency of estimators of the overall population parameters by the choice of strata that is homogenous over the sampling which are within stratification which makes the survey easier to administer operationally.

4.3 Target population and sample size

The sample used in this study was drawn from Ndhiwa sub – county in Homa - bay county .The research targeted the 2013 KCSE candidates. The total number of students in the target population were 1715, out of these 319 were from county schools whole 1396 were from sub – county schools. The standard deviation of the county schools and sub – county schools were 10.27 and 6.67 respectively.

In determining the sample size for the population, we assumed a level of significance (α) of 5%; a level of confidence of 95% and the degree of variability to be 50%.

Using the formula:

$n = \{ Z^2pq + ME^2\} / \{ME^2 + Z^2pq/N\}$

Where: n = the sample size

ME = Marginal error

N = Total population size

P, q= Degree of variability

$n = (1.96^2 * 0.5 * 0.5 + 0.05^2)/\{0.05^2 + [(1.96^2 * 0.5 * 0.5)/1715\}$

$= 315$

In our research we sampled a total of 460 candidates. We next proportionately allocate this sample to the two strata county schools and sub – county schools.

Based on Neyman allocation, the best sample for stratum h is given by

$n_h = n(N_h * \sigma_h)/ [\sum(N_i \sigma_i)]$

Where: n_h = Sample size stratum h

n = Total sample size

N_h = Population size stratum h

σ_h = Standard deviation of stratum h

$n_{cs} = 460 (319 * 10.27)/ \{(10.27 * 319) + (6.67 * 1396)\}$

$=120$

$n_{scs} = 460 - 120$

$= 340$

4.4 Data Collection

Secondary data was obtained in documented materials from various secondary schools and the sub –county education offices.

CHAPTER FIVE: DATA ANALYSIS AND PRESENTATION

5.1 INTRODUCTION

This chapter reports study findings by representing a comprehensive analysis of descriptive statistics, multiple linear regression analysis properties and the multiple regression analysis of the KCSE score with the KCPE aggregate marks, Mathematics scores, English score as the predictor variables.

5.2 Descriptive Statistics

5.2.1 Correlation Analysis

To identify if there is a correlation between KCSE scores and the KCPE aggregate marks, Mathematics scores and the English scores variable the study used correlation co –efficient. The study below gives summary of the correlation co – efficient.

VARIABLE	KCSE	KCPE	MATHEMATICS	ENGLISH
KCSE	1.000	0.742	0.601	0.604
KCPE	0.742	1.000	0.787	0.787
MATHEMATICS	0.601	0.787	1.000	0.664
ENGLISH	0.604	0.787	0.664	1.000

Table 1: correlation co- efficient table

The result shown in table 1 shows that there is relatively weak positive correlation between KCSE scores and the English scores; there is a strong correlation between KCSE and the Mathematics scores but a very strong correlation between the KCSE and the aggregate KCPE marks. From the results the study rejects the null hypothesis of no correlation thus $r \neq 0$ at 5% significance level.

Although the correlation co – efficient measures the co variability of variables it does not necessarily imply any functional relationship between the variables concerned. In addition it does not establish and/or prove any causal relationship between the variables.

5.2.2 MULTICOLLINEARITY

Multicollinearity is a term reserved to describe the case when the intercorrelation of the predictor variables is high. Multicollinearity occurs when two or more variables contain strongly redundant information. If two or more variables are multicollinear then there is not enough distinct information in these variables for the multiple regression to operate correctly.

A multiple regression with two or more independent variables that measure essentially the same thing will produce errant results. Variables which are multicollinear are indistinguishable in the regression line i.e. multicollinearity mathematically corrupts the linear model. The following signs indicate multicollinearity:

(i) High collinearity between pairs of predictor variables.

(ii) Regression co – efficient whose signs or magnitude do not make good physical sense.

(iii) Statistically non – significant regression co – efficient on important predictor variables.

(iv) Extreme sensitivity of sign or magnitude of regression co – efficient to insertion or addition of predictor variables.

When there is a perfect linear relationship among predictor variables the estimate for regression cannot be uniquely computed implying the variables are near perfect linear combination of one another. As the degree of multicollinearity increases, the regression model estimates of the co – efficient becomes unstable and the standard error for the co –efficient gets widely inflated.

To establish multicollinearity between variables we compute the Variance Inflation Factor (VIF), correlation co - efficient and the tolerance for each of the predictor variables.

VIF =1/Tolerance

If the value of the VIF is greater than 10 then we may suspect multicollinearity of variables hence further investigation on the data may be done otherwise the variables not multicollinear

The tolerance value on the other hand shows the percentage of variance in the predictor variable that cannot be accounted for by the other predictor variables. If the value of the tolerance is less than 0.10 then we may suspect multicollinearity and hence further investigation of data may be necessary otherwise the variables are not multicollinear.

The values of the correlation co – efficient from the correlation matrix which compares independent variables with each other can also be used to check multicollinearity. If the correlation coefficients are above 0.80 then variables are multicollinear.

The table 2 below gives summary for the results of multicollinearity indicators of the study

MULTICOLLINEARITY STATISTICS

MODEL	TOLERANCE	VIF
CONSTANT		
KCPE	0.255	3.921
MATHEMATICS	0.375	2.666
ENGLISH	0.375	2.666

Table 2: Collinearity statistics

From the table 2, the tolerance level of KCPE (0.255); MATHEMATICS (0.375) and ENGLISH (0.375) are acceptable subject to the condition that they are more than 0.10. Also the VIF of KCPE (3.921), MATHEMATICS (2.666) and ENGLISH (2.666) are acceptable subject to the condition that they are less than 10.

Based on the above VIF and tolerance level of the predictor variables we can conclude that there is no perfect linear relationship among the KCPE scores, Mathematics scores and the English scores. These predictor variables are not perfect linear combination of one another and hence the intercorrelation between the KCPE, Mathematics, and English are not high thus they are not multicollinear.

Considering the correlation co- efficient in table1 i.e. KCPE (0.742), MATHEMATICS (0.601), and ENGLISH (0.604) all are well below 0.80 a proof that the predictor variables are not multicollinear.

5.2.3 R – SQUARE (R^2)

The R – square value indicates the percentage of variation in the KCSE and KCPE, Mathematics, and English that is explained by the model. The value of R^2 lies between 0 and 1. A value close to zero indicates little association between the set of independent and dependent variables while a value near 1 means a strong association.

The number of independent variables in the multiple regression equation makes the coefficient of determination larger. Each new independent variable causes the prediction to be more accurate. The R^2 increases only because of the total number of independent variables and not because of the added independent variable is a good predictor of the dependent variable. To balance the effect that the number of independent variable has on the coefficient of multiple determination we find the adjusted co efficient of determination given by:

$$\text{Adjusted } R^2 = 1 - \{[SSE/(n-k-1)] / [SST/(n-1)]\}$$

The table below shows the R – Square parameters of the study:

Model Summary

Model	R	R – Square	Adjusted R – Square	Standard error deviation of estimate
1	0.743	0.553	0.550	0.8400

Table 3: R – Square table.

In our research the value of R – Square adjusted was 0.550. This shows that 55% of variations in the 2013 KCSE results in Ndhiwa sub - county were explained by the KCPE, Mathematics, and English scores. The remaining 45% are explained by other variables not included in the model.
It should be noted that the strength of relationship between the predictor variables may be driven by other variables "lurking" in the background which are related to the independent variables. This makes it hard to reliably find causal relationship. The association established would as a result of other variables in the background. Hence we can conclude that using KCPE, Mathematics, and English scores as predictor variables the model accounts for 55% variability in the KCSE scores.

5.2.4 REGRESSION COEFFICIENTS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	-33.569	3.838		-8.745	.000
	KCPE	.273	.025	.675	10.883	.000
	MAT	.040	.055	.037	.731	.465
	ENG	.054	.057	.048	.940	.348

a. Dependent Variable: KCSE

Table 4: Regression co – efficient table

From the regression co – efficient table 4 above we have ignored the intercept in the model because KCPE, Mathematics, and the English scores do not have zero values. It`s also unlikely that a student who had a total marks of zero in the KCPE would find his way into one of the secondary schools in the sub – county. Since the KCPE, Mathematics, and the English scores are measured in similar scale we compared their regression co – efficient. From the data output the KCPE scores mattered more in predicting the KCSE than Mathematics and English. However the English scores were slightly better than the Mathematics scores in the same predictions. This could be attributed to the fact that English is the medium and having strength in the subject has a ripple effect in the other subjects. We may say statistically adjusting for the effects of KCPE score, Mathematics scores, and English scores each one unit increase in the KCPE scores is associated with 0.273 increase in the KCSE scores; each one unit increase in the Mathematics scores is associated with 0.040 increase in the KCSE scores also each one unit increase in the English scores is associated with 0.054 increase in the KCSE scores.

The beta coefficient (beta weight) which measures the associations in standard deviations units also support the idea that KCPE marks is a better predictor than the Mathematics score and the English score i.e. KCPE (0.025), Mathematics (0.055) and English (0.057).

5.2.5 SCATTER PLOTS

This is a scatter plot of the residuals against the predicted values. The residuals should lie around zero with the degree of scatter not varying systematically with the size of the predicted values. If the values between the observed values and predicted values of the dependent variable are computed we obtain unstandardised residuals which are scale variant, we can make the dependent variable scale invariant by standardizing them.

However the variances of both, the raw residuals and the standardized residuals vary with the values of the explanatory and in that sense these residuals are not optimal for examining the error distribution (which are not constant variances) studentised residuals (*SRESID) are corrected for this effect of the explanatory variables and should if the model is correct arise from the studentised normal distribution, it is this residuals that we used.

The residual plot was generated by plotting *SRESID on the y – axis and *ZRESID (predicted values of the standardized to mean zero and standard deviation one) on the x – axis to obtain the graph below.

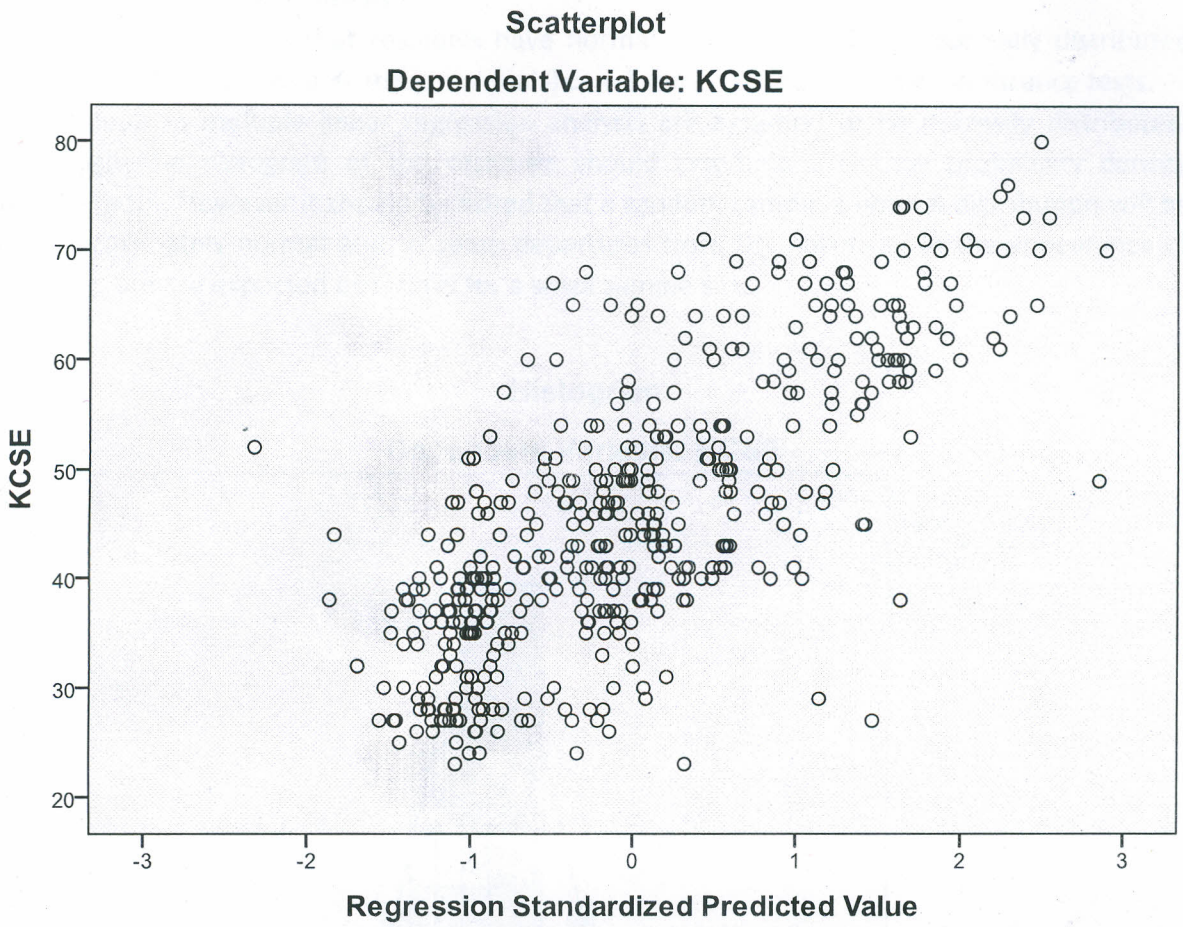

Figure1: figure of scatter plots

The scatter plot in figure 1 above reveals a linear relationship between the KCSE scores and the standardised predicted value of the KCSE scores. For a given value of Mathematics, KCPE aggregate score and the English scores the predicted value of the KCSE score will fall on the line. The plot also further reveals that the variation in the KCSE score about the predicted value is about (+ or – 10) units regardless of the value of X. Statistically this is referred to as homoscedacity. Violation of homoscedacity leads to parameter estimates with inflated variances. However it should be noted that when the plots of the residuals seems to deviate so much from the normal more formal tests for heteroscedacity should be formed. Possible tests are the Fold feld – Quandt test.

5.2.6 Histogram of the residuals

The regression assumes that residuals have normal distribution. Non – normally distributed variables (highly skewed or Kurtotic residuals) can distort relationships and significance tests.

The residuals in multiple linear regression analysis are assumed to be normally distributed. Accordingly the histogram of the residuals should resemble a normal probability density function (pdf). However it should be noted that a random sample a normal distribution will be only approximately normal and so some departures from the normality in the appearance of the histogram are expected especially for a small sample size.

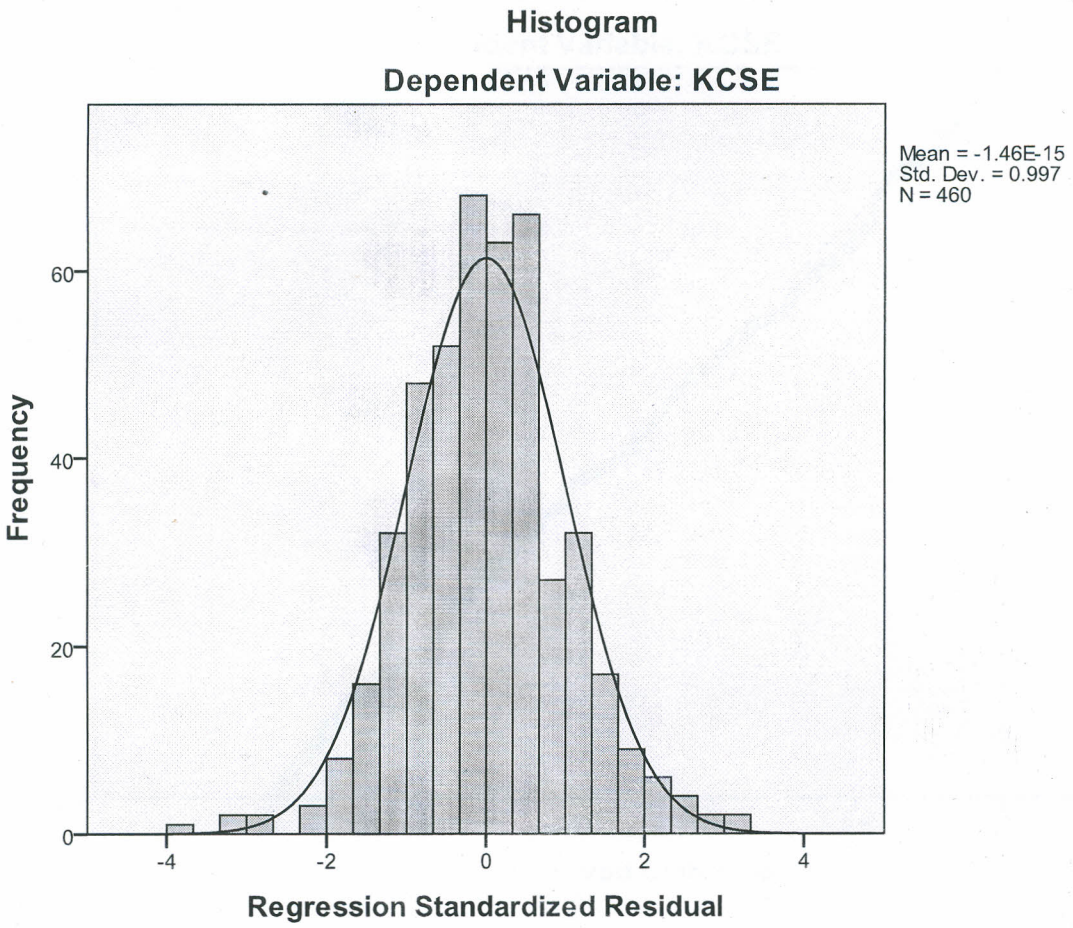

Figure 2: Histogram of residuals

From figure 2 above the normality assumption is not violated since the histogram of the residuals is consistent with the assumption of normality i.e. the mean is -1.46E-15 which is approximately zero and the variance is also 0.994.

5.2.7 Normal probability plot

This is a graphical technique for assessing whether or not a data set is approximately normally distributed. The observed data is plotted against the theoretical normal distribution in such a way that the lines should form an approximately a straight line. Departures from this straight line indicate departures from the normality.

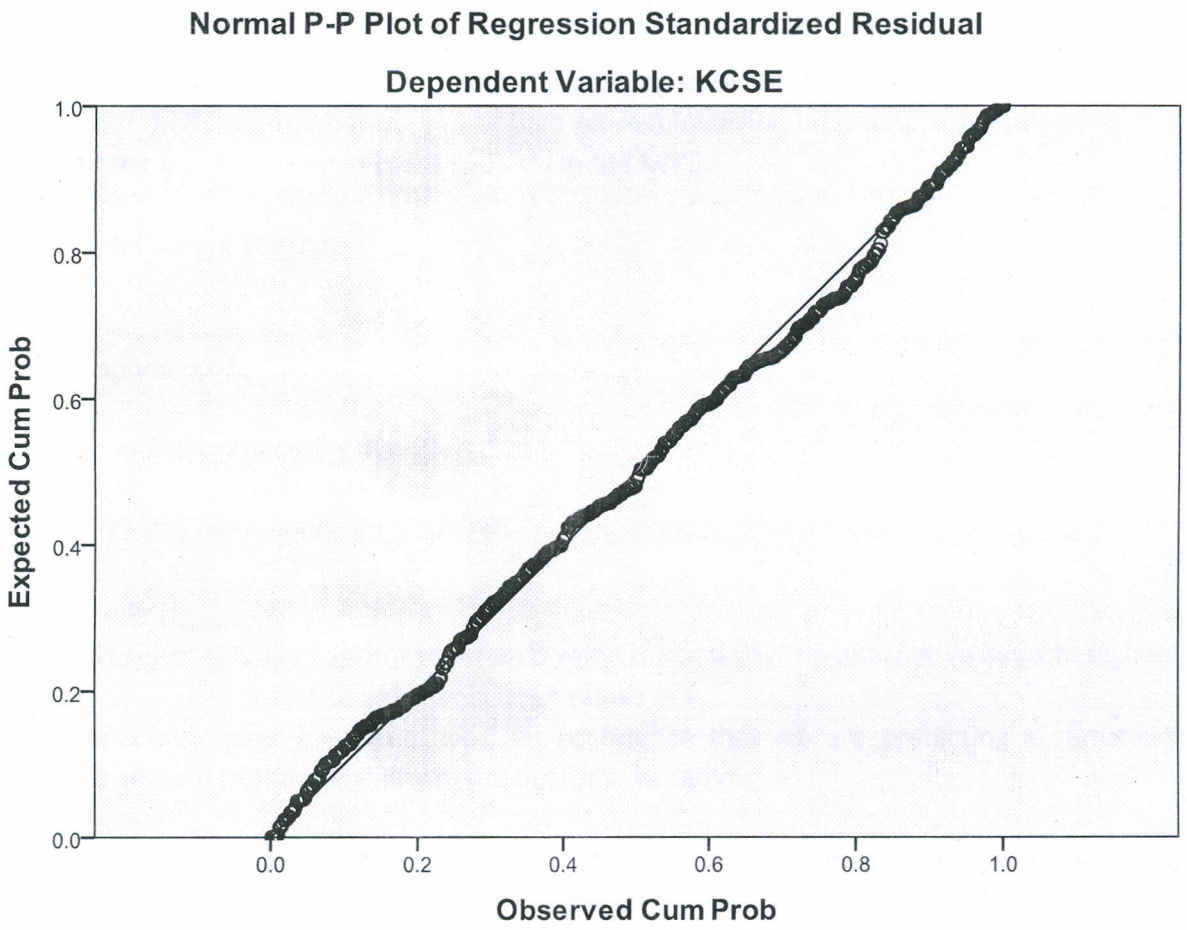

Figure 3: Normal p – p plot

From figure 3 above the points on the plot form a nearly linear pattern which indicates that the normal distribution is a good model for this data set.

5.2.8 THE F – TEST

The F – statistic generally helps measure how far our regression prediction is from zero or how well our model is doing in predicting the KCSE scores. The F – test gives whether the regression equation as a whole is useful in making predictions, that is whether the variables X_1,X_2 and X_3 taken together as a group are useful in predicting y (the KCSE scores). This testing procedure looks at the overall test of significance which helps us to determine whether or not our regression is worth anything. It helps in revealing whether our model is predicting the outcome variable better with the information from the explanatory variables.

In our research we are interested in the relationship between the KCSE score to the specific subjects (i.e. KCPE, Mathematics, and English scores) involving interactions between the KCSE score and the grades scored in these subjects in the KCPE.

In our model we are testing:

H_0: $\beta_1 = \beta_2 = \ldots\ldots\ldots\ldots = \beta_k$.
H_1: At least one $\beta_j \neq 0$.

Under the null hypothesis f is distributed as:
$f \sim f_{p,n-p-1}$
Generally $f > 4$ is very significant (reject the null hypothesis). The p – value for the f – test is

P – Value = $Pr\ (f_{p,\ n-p-1} > f)$
We usually reject the null hypothesis when p value is less 0.05. The alternative hypothesis holds for either very small values or extremely large values of f.
If we find a significant f – value, we gain confidence that we are predicting a statistically significant amount of the variability in the outcome variable.

The information provided by the regression identities for the sum of squares and the degrees of freedom and the F – statistic are usually summarized in a table called analysis of variance (ANOVA). The table 5 below shows the ANOVA table for our data.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	39738.924	3	13246.308	187.711	.000[a]
	Residual	32178.807	456	70.568		
	Total	71917.730	459			

a. Predictors: (Constant), ENG, MAT, KCPE

b. Dependent Variable: KCSE

Table 5: ANOVA Table

The regression identity for the sum of squares tells us that

SST = SSR + SSE

Using these facts and a bit of algebra it can be shown that the f – statistic can be computed from the value of R^2 by using the following formula:

$F = \{R^2/k\} / \{(1 - R^2)(n-k-1)\}$

If R^2 is close to zero, the value of F will be small. This case corresponds to the situation in which little variation in the observed values of the response variable can be accounted for by the multiple linear regression equation in the predictor variables. If the value of R^2 is close to 1, the denominator term involving $1 - R^2$ will be close to zero and the value of F will be large. The case that R^2 is close to 1 corresponds to a large proportion of the variation in the observed values of the response variable being accounted for by the multiple linear regression equation.

5.2.9 The t – test

The t – test assists us to make inferences concerning the utility of a particular predictor variable. The t – test helps us to decide whether a particular variable say mathematics, English or KCPE is useful for predicting the KCSE scores. It helps in revealing whether our model is predicting the outcome variable better with the information from each explanatory variable X_1, X_2, and X_3.

In our research we are interested in the relationship between the KCSE score to the specific subjects (i.e. KCPE, Mathematics and English).

We therefore perform the hypothesis test:

H_O: $\beta_i=0$

H_1:$\beta_i\neq0$

Rejection of the null hypothesis indicates the X_i is useful as a predictor of y (the KCSE scores) and that it may be worthwhile to do regression analysis with the variable X_i omitted.

We use the sample regression co – efficient b_i to estimate β_i and b_i will be the basis for our test statistic of the null hypothesis H_0: $\beta_i = 0$.

The value of the t – statistic is given by t

$t = b_i/s_{bi}$

Where: b_i – The co – efficient X_i predictor variable.

s_{bi} – Standard error of the X_i predictor variable.

In order to make a decision on whether to reject the null hypothesis or not we performed our hypothesis test at 5% significance level and so $\alpha= 0.05$.

The t – statistic has d.f = n-(k+1), where n is the total number of observations and k is the number of predictor variables and the p – value is obtained from the spss output.

If $p \leq \alpha$ reject H_o otherwise do not reject the H_o.

The table below shows the predictor variables co – efficient, their standard error and the corresponding p – values for our data.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	-33.569	3.838		-8.745	.000
	KCPE	.273	.025	.675	10.883	.000
	MAT	.040	.055	.037	.731	.465
	ENG	.054	.057	.048	.940	.348

Table 6: Table of co - efficient

From table 6 the p – value for the KCPE is less than our significance level i.e. 0.000≤ 0.05 hence we reject the null hypothesis. The data provides very strong evidence against the null hypothesis. The test results are statistically significant at 5% significance level; the data provides sufficient evidence to conclude that the regression co – efficient β_1 of the population is not zero. Hence in conjunction with Mathematics and English scores KCPE is a useful predictor of the KCSE scores.

Also the p – value for the Mathematics is more than our significance level i.e. 0.465 ≥0.05 hence we accept the null hypothesis. The test results are not statistically significant at 5% level of significance; the data provides strong evidence to conclude that Mathematics scores do not contribute significantly to the KCSE scores.

On the other hand the p – value for English also is more than our level of significance i.e. 0.348≥0.05 hence we accept the null hypothesis. We therefore conclude that English does not contribute significantly to performance in the KCSE.

5.2.10 CONCLUSION AND RECOMMENDATIONS

Mathematical models are important tools in scientific research because of their ability to simulate, more or less exactly, a particular physical situation. They provide the researcher with means for prediction and decision – making. In this research we have managed to develop the following multiple linear regression model

$Y = -33.569 + 0.273X_{1i} + 0.040X_{2i} + 0.054X_{3i}$

Where X_{1i} = is the i^{th} KCPE scores.

X_{2i} = is the i^{th} mathematics scores.

X_{3i} = is the i^{th} English score.

According to the data both the KCPE, Mathematics and English score were statistically significant in the regression model at 0.05 significance level. Therefore it can be concluded that these entry behavior of students to the secondary education contribute much to their success at the secondary schools. However it should be noted that KCPE aggregate marks is a better predictor of the KCSE scores followed by mathematics then lastly English. Also other factors such as school tradition, administration, infrastructures and the experience and qualifications of teachers may contribute to variation in performance at the KCSE level.

From the result one percent increase in KCPE aggregate score increases the KCSE scores by 0.273.It was further concluded that one percent increase in the mathematics score increases the KCSE scores by 0.040 in the secondary school and that one percent increase in English scores at the primary school increases the KCSE scores by 0.054.

Therefore based on the results there is a good reason to incorporate the Mathematics and the English scores in the form one selection since their combined effect is significant in the prediction of the KCSE scores as a tested by the F – test. .

There were a number of issues that this study was unable to address due to its scope. In view of this the following were recommended for further research in our multiple linear regression model.

(i) KCSE results depend on several factors such as school administration, teacher student ratio, School infrastructure and resources, gender and so on. A multiple linear regression model that takes into consideration these factors should be considered.

(ii) Future models should also take into consideration environmental problems and home backgrounds of the students.

(iii) Since English is the language of instruction both at the primary and secondary, a lot of emphasis should be given to it so that the ripple effect can be felt in the other subjects.