MODELING THE PERFORMANCE OF SCIENCE SUBJECTS FOR SECONDARY SCHOOLS IN KENYA USING PARTIAL LEAST SQUARES REGRESSION

BY

MWITI MIRRIAM KARWITHA

A RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTFOR THE DEGREE OF MASTER OF SCIENCE IN SOCIAL STATISTICS OF THE UNIVERSITY OF NAIROBI

NOVEMBER 2016

DECLARATION

I hereby declare that this research paper is my original work, which has not been undertaken anywhere else in any university for the award of a degree.

Name Sign Date

Mwiti Mirriam Karwitha …… …….

REG NO.: I56/75438/2014

APPROVAL

This paper has been submitted for examination with my approval as the university supervisor

Supervisor Sign Date

Dr. Nelson Owuor…..…… ……..

ACKNOWLEDGEMENT

I am very thankful to my supervisor Dr. Nelson Owuor for the great supervision and guidance. My sincere appreciation also goes to all the school of mathematics lecturers who assisted me in my course work. I would also like to thank my classmates for their assistance and support. Special thanks to my classmate Joseph Kuria and Victor Otieno for their efforts and assistance. My sincere appreciation goes to my family for their great support. I want to thank the Almighty God for His sufficient grace that has made this possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND ACRONYMS

INSET	In service Education and Training
KCSE	Kenya Certificate of Secondary Education
KNEC	Kenya National Examination Council
MOE	Ministry of Education
PLSR	Partial Least Square Regression
SMASSE	Strengthening of Mathematics and Science in Secondary Education

ABSTRACT

Students' performance in sciences is closely associated with the scientific and technological innovations worldwide. The Government of Kenya recognizes the important role science must play in achieving "Vision 2030" and invest resources in raising the quality of teaching mathematic science and technology. English, a second language in Kenya communities is the language of instruction and assessment in schools. As the social sciences develop, hypothesized relationships become increasingly more complex, therefore the need to use more versatile models. Partial Least Square Regression is one of those models. The purpose of the study was to investigate the effect of learners' achievement in literacy on the performance in science subjects. From the study it can be concluded that both mathematics and languages contribute highly in the development of science process skills. The study recommends that other than focusing on mathematics and sciences, they should also focus on the development of literacy skills in English and Kiswahili. Policy-makers should also consider planning for capacity development trainings for English and Kiswahili.

CHAPTER ONE

INTRODUCTION

1.0 Background to the Study

Education is considered a key factor of social economic development thus nations all over the world spend fortunes to enhance the education process. In addition, education improves the productive capacity of societies and their institutions. It also strengthens the value and efficiency of labour provided by the learned and empowered workers. This is in agreement with Mwaura (2010) who says "as technology advances, new methods of production depend on a well-trained and intellectually flexible labour force."

Mathematics has a direct correlation with other subjects, mostly technical and sciences making it an important subject that is included in the curriculum worldwide. In Kenya it cuts across all primary and secondary school as a compulsory subject. According to Tshabalala and Ncube, (2013), "mathematics as a vital tool for scientific, technological and economic expansion of any nation". Davies and Hersh, (2012) are of the same opinion whereby they view mathematics as an essential subject that prepares the students for the future irrespective of the career path they choose. Mathematics is therefore intimately connected to everyone's long life planning, implying that both education and human life cannot function successfully without it.

On the other hand, sciences are subjects that are of equal importance as mathematics since they are seen by society as the foundation of technological knowledge that is fundamental in social-economic development of the nation. The importance of science subjects is also recognized internationally not only for economic welfare of nations, but because a scientifically literate community is required. Mususya (1992) views science as a subject

that equips learners with manipulative skills that are important for industrial development and hence economic growth. He also adds that sciences are used as a basic entry requirement into the prestigious courses such as medicine, architecture and engineering

Equally, Kenya recognizes the importance of science and mathematics in the realization of its vision. This is reflected in the amount of resources both human and otherwise that are channeled towards enhancing the teaching and learning of science and mathematics subjects at all levels of the education system. At secondary school level, a number of intervention strategies have been put in place to ensure that teaching and learning of these subjects is as effective as possible. Apart from providing trained teachers to handle these subjects, the government has institutionalized In-service Education and Training (INSET) that involve teachers who teach mathematics and sciences under Strengthening of Mathematics and Science in Secondary Education (SMASSE) programme. A substantial quantity of funds from Ministry of Education's budget is contributed towards this programme (MoE, 2005). Schools on the other hand have been charged with the responsibility of providing learning resources through their BOM and PA committees.

Table 1.1: Performance in sciences compared to other subjects nationally between 2010 and 2015

Subject	2010 Overall mean	2011 Overall mean	2012 overall mean	2013 Overall mean	2014 Overall mean	2015 Overall mean
ENGLISH	39.26	36.74	38.13	35.23	47.68	40.29
KISWAHILI	44.34	49.01	36.32	39.91	47.68	47.88
MATHEMATICS	19.17	21.00	25.30	25.10	24.02	26.88
BIOLOGY	26.71	31.72	25.38	28.70	31.83	34.8
PHYSICS	31.50	32.94	32.53	36.87	38.84	43.68
CHEMISTRY	22.89	23.40	27.72	25.45	32.16	34.36
HISTORY	41.73	38.45	37.14	41.78	53.83	51.71
GEOGRAPHY	33.86	38.15	43.09	41.02	44.02	43.92
CRE	46.05	49.38	44.34	51.93	53.15	52.48
AGRICULTURE	31.25	34.26	32.03	31.94	40.82	44.81
BUSINESS	37.28	42.61	51.00	53.64	46.82	43.76

Source: KNEC (2015)

Table 1.1 indicates that on average, mathematics and sciences are performed more poorly than other subjects. In 2010, the overall performance of the students in sciences was generally low compared to other subjects. Chemistry for example with a mean of 22.87% was only better than mathematics with mean of 19.17%.In 2011, the performance in sciences was still low with mathematic being the worst with a mean of 21.0% then followed by chemistry having a mean of 23.4% this was only slightly better than Biology and Physics with a mean of 31.72% and 32.94% respectively .In 2012, the trend was such that

Biology with mean of 25.38% was only slightly better than mathematics with a mean of25.30%, chemistry and physics were lagging behind the rest of the non-science subjects. In 2013 and 2014, the trend was similar to the one in previous years where chemistry was performed slightly higher than mathematics while biology and physics were lagging behind the rest of the subjects. It is important to note that in every year that English and Kiswahili were performed well, there was a slight improvement on the science subjects.

1.1 Problem statement

Good level of understanding of mathematics and sciences is an indispensable component for a successful, professional and social life. Therefore, the failure rate in sciences in KCSE examination (which is the minimum level of basic skills in mathematics and science subjects) being above 50% for young Kenyans is alarming. Unlike many other countries where by policies are directed on addressing the low achievement in literacy, Kenya's policies are mainly on addressing low achievement in mathematics and sciences with little being done on literacy. Buindi (2013) also notes that despite the importance given to sciences, there has always been poor performance in the sciences at national examinations compared to other subjects. Establishing whether their exist a relationship between achievements in literacy, achievement in numeracy skills and processing skills could lead to a better understanding of the poor performance in science subjects in Kenya.

1.2 Objectives of the study

The main objective of the study is to investigate the effect of learners' achievement in literacy on the performance in science subjects with specific objectives being:

- To establish the relationship between achievement in language, achievement in mathematics and achievement in science subjects.
- To model the performance in sciences given the grades in English, Kiswahili and Mathematics using partial least squares regression (PLS-regression)

1.3 Justification of the study

While studies like Muola et al(2013), kipkorir(2013), Daniel and Sifuna (2007) ,and Andile and Moses (2006), have studied performance in science and mathematics using explorative and descriptive statistics; such studies do not specifically consider how achievement of literacy affect the achievement of the science and mathematics subjects .Therefore the approach taken in this project is to use PLS regression as the analytical tool rather than statistics methods used on the previous studies to establish literacy as factor affecting achievement in science subjects. Furthermore, if the relationship between languages, mathematics and sciences is established, it will help to advice education policy makers in lying more emphasis in achievement on literacy skills as an effective measure in achieving good performance in sciences.

1.4 Organization of the study

The rest of the paper is organized as follows: chapter two reviews the literature on PLS regression modeling technique and the literature on the influence of language of instruction on performance in mathematics and science subjects. The methodology used to analyze data is presented in chapter three. Chapter four provides the results and data analysis; chapter five is on the conclusion and recommendations of the study.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

In this chapter a review of the literature on both the origin and application of PLS-R and literature related to achievement in mathematics and sciences as well as the proficiency in the English language in education is presented.

2.2 The background of PLS

Partial Least Squares (PLS) is a technique that is used to build a model where there is a huge number of correlated explanatory variables. The origin of this technique goes way back in 1966 when Herman Wold presented two iterative procedures using least squares (LS) estimation for both single and multicomponent models and canonical correlation. In 1973, these two iterative procedures gave way to (NIPALS) algorithim with which Wold presented a way to calculate principle components and canonical correlations with an iterative sequence of simple and multiple regressions using ordinary least squares (OLS) respectively. According to Fornell (1994), both of Wold's iterative algorithms were later followed by the PLS algorithm which was originally known as NIPAL for least squares estimation of path models with latent variables. Geladi (1988) summarizes it all by saying, "Herman Wold gives the end of 1977 as the birth date of PLS".

2.2.1 The origin of PLS-Regression

Having showed how to use the iterative sequence of both simple and multiple OLS regression to compute principle component and canonical correlations, Wold continued to apply the technique to new problems and fields. In the 1980's the interest of application of PLS in research shifted from social sciences to chemo metrics. This change was initiated

by Svante Wold (son of Herman Wold), whom together with Harald Martens modified NIPALS to solve the problem of multicollinerity in linear regression. Svante and Harald developed yet another branch of the PLS technique known as PLS regression which was initially significant in analytical chemistry.

2.2.2 Application of PLS-Regression

PLS-Regression has not only provided a solution to the problem of multicollinearity in regression models, it also solves the problem that arises when the predictor variables are larger than the observed variables.PLS has become a prevailing analytical tool in predictive regression modeling with multiple and diverse applications. Although it is widely used in chemometrics, PLS regression has also importance in other field of study like bioinformatics (Gersende and Sophie, 2004), sensometrics, neuroscience (Marten and Naes 1989) and anthropology. Due to the problem of small sample size, correlated and high number of predictors, and high noise to signal relationships, ecologist have opted to use PLSR as an alternative to current regression methods used in ecology (Luis, Ismael and Oscar, 2009). McIntosh, Bookstein, and Grady (1995) introduced PLS method as a new tool for functional neuro-image analysis because of its exclusive way of generating spatial patterns of brain activity therefore explaining the relationship between image pixels and task or behavior.

Many Econometric models which include time series data have a multcollinearity problem because economic variables usually have correlations with each other. In such situations, decomposition by PLSR is ideal tool since it is designed to deal with this condition (Ozlem and Gulder, 2012). PLS has also received substantial attention in other fields such as

computational biology (Tan et al., 2004), geosciences for purposes of paleoclimate reconstructions (Kalela-Brundin, 1999) and statistical prediction (McIntosh et al., 2005)

2.3 The theory of language in learning science and mathematics.

Language is the vehicle of discretion that transmits the intended message to the receiver. It is also viewed as a means through which thought is organized, refined and expressed. Therefore, language helps in formation of concepts, analysis of complex ideas and focusing attention on ideas which would otherwise be difficult to comprehend in absence of language. Smith and Ennis (1961) viewed the use of language in school as both an instrument and the vehicle of interaction between the teacher and the student since the conduct of classroom instruction is inescapably involved in the use of interpretation of language which could either be inform of writing, print or verbal. Based on this argument one can therefore conclude that for a student to achieve, he must have a good command in the language of teaching and learning. Whorf (1956) suggested that language defines and determines how we think.

English is used as the medium of instruction in teaching and assessing achievement of both mathematics and science subjects in Kenya. When the language of tutoring is English, learning the other subjects like maths, biology chemistry and physics raises some issues for learners whose English is a second language. Learning a second language becomes difficult mainly when the language is first learned in the classroom. For learners to effectively learn mathematics and sciences, different linguistic skills are required, which a second language learner may not have mastered.

Language allows students to participate in classroom activities thereby accessing the accurate subject matter as defined in the syllabus. Lee et al. (2013) suggest that "teachers promote academic language acquisition by supporting students' ability to do things with language, engaging them in decisive activities and providing them with opportunities for language use". Barber (2001) says that ways in which learners make sense of science practical is influenced by the principles, attitude, experience, communication patterns as well as teaching styles. Therefore, in order to understand concepts in science, students must learn how scientific knowledge is constructed, represented and communicated.

Thinking scientifically involves the appropriation of the ways scientists use language. Gee (2005) and Lemke (1990) have both suggested that the scientific thinking of a student is intertwined language. Halliday and Martin (1993) added that scientific communication consists of variety of genres which consist of different patterns of linguistic features that are arranged in such a way that certain aspects of scientific knowledge and reasoning can be communicated efficiently. Therefore, scientific researchers acknowledges the important of language skills in the inquiry processes. Science teachers therefore need to scrutinize the syllabus and resources for the academic language, add in literacy strategies into teaching and give chances to students to exercise their language fluency through reading writing and speaking to ease comprehension.

On a review report on numeracy in National Numeracy Review Report, the importance of language in learning mathematics is recognized. In the report it was suggested that language and mathematics literacy should be explicitly taught by all the teachers who teach mathematics in acknowledgment that language form a barrier to understanding of mathematics concepts. In the model proposed by Clark in 1995 to represent various roles

that would be played by language in teaching and learning of mathematics, he views concepts as a result of learners' knowledge, with language steering the learners' conceptual growth through instruction and discussions. Thorndike (1912) concludes by noting that the measure of the ability in arithmetic, is a measure of both the sheer of mathematic and knowledge and the acquaintance with language.

2.4 Performance in mathematics and sciences

Much of the research done on the poor performance in science has revolved around the attitude towards the subjects, availability of teaching resources and background of student which many talks of the role played by the parents on their child's education. In his study on factors contributing to poor performance of science subjects in public school in Acare, Nigeria James (2014) found out that the most common factors included negative attitude toward the science subjects among students and lack of resources such as well-equipped laboratories and textbooks. Karue and Amukowa (2013) recommended that if instructional materials, library, laboratory and other physical facilities were provided, and the teacher student ratio was reduced, that would be one of the ways of improving performance in mathematics and sciences.

Andile and Moses (2006) used a non-experimental, exploratory and descriptive method to study the factors that are linked with high school leaner's poor performance with attention being given to mathematics and physical science in District 3 of Tshwane North. They found out that the teaching strategy, content knowledge, motivation, and incomplete syllabus coverage had a direct influence on the poor performance while the role of parents in child education and the use of language as well as its understanding had indirect influence on the performance of the two subjects.

A study by Daniel and Sifuna (2007), evaluated the value of the SMASSE and that of SbTD programmes on classroom relations in all schools in Kenya. They found out that, while teachers evaluated the though the two programmes were helpful in exposing them to a student-centred approach, their classroom practices did not reflect it since approached largely dominated by the teacher. This is partly associated with pressure of covering the syllabuses in preparation of the final exams, high population of pupils in classes, and the use of English as a second language.

2.5 Overview of literature

Based on the above discussion it is important to note despite the recognition of the importance of language in learning both mathematics and science subjects, research on the connection between English verbal communication and learning concepts along with skills obtained from science and mathematics is scarce. Although studies like Daniel and Sifuna (2007) and Andile and Moses (2006) have pointed out that English could be a determinant of poor performance in the science subjects much attention was given to other factors in the study. Descriptive surveys are the main statistical tool that has been used in the previous studies on poor performance in sciences. Therefore, further studies in this area using different statistical techniques may be value adding especially with the focus being language as the determinant.

CHAPTER THREE

METHODOLGY

3.1 Introduction

Chapter three presents the type of data used in the study, the procedures and the method used in the analysis so as to achieve the objectives of the study.

3.2 Research Data

The study will use the Kenya Certificate of Secondary Education (KCSE) 2014 results from Kenya National Examination Council (KNEC). KNEC is the body responsible in conducting public academic technical and other examinations at basic and tertiary levels. KCSE examination is done after completion of the four years of study in Kenyan secondary schools. The study population was all the 438660 candidates who received their 2014 KCSE results while the target population was all the 65535 candidates who undertook all the three science subjects namely, Biology Chemistry and physics. Data analysis was conducted using the PLS-Regression technique to establish the relationship between mathematics, languages and the science subject and predict the achievement of the science subjects given the performance in mathematics and the languages.

3.3 Definition of the variable

The independent variables are the influencing factors that determine the achievement of the processing skills. The factors include the numeracy skills which obtained as grades in Mathematics, the literacy skills obtained as grades in Kiswahili and English, gender of the student and the school type. Processing skills is the dependent variable which is a multinomial response since it has three categories which including the grades in Biology Chemistry and Physics.

3.4 Assumptions underlying PLS-Regression

In PLS modeling, it is assumed that there exist some causal variables which are known as latent variables (LV's) that actually influence the system that is being investigated. The exact number the of these causal variable is generally estimated in the PLSR process since they are not known, making it one of the aim of the analysis. Both the response and predict variables are assumed not to be independent because they are assumed to be manifest of these causal latent variables. PLSR is theoretically based on one of Taylor expansions. It is assumed that multi-dimensional function will generate data for two data matrices (X and Y). A latent variable model is then generated by expanding the function using Taylor expansion method. PLSR also assumes that the data is homogeneous implying that the means of influence of X on Y should be the similar.

3.4 PLS-R Model

Partial Least Squares methods relate linearly two data matrices which are put in blocks, by means of an underlying latent variables model generally given as

$$Y = TB + F^{*}; \qquad (1)$$

With T denoting the latent variables matrix which is also known as score matrix and F denoting the residual matrix. Given below is the structure of the data with the X-block and the Y -block in matrix form to set the general frame

$$X_{(n\times N)} = \begin{pmatrix} x11 & x12 & \cdots & \cdots & x1N \\ x21 & x22 & \cdots & \cdots & x2N \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ xn1 & xn2 & \cdots & \cdots & xnN \end{pmatrix} y_{(n\times M)} = \begin{pmatrix} y11 & \cdots & y1M \\ y21 & \cdots & y2qM \\ \vdots & \vdots & \vdots \\ yn1 & \cdots & ynM \end{pmatrix}$$

PLS models the relations of two blocks of variables by decomposing the two blocks of zero mean to form

$$x = tp^T + E$$
$$y = uq^T + F \qquad (2)$$

Where by the t and u are matrices of the extracted score vectors (components, latent vectors), p and q represent matrices of loadings and while E and F are the matrices of residuals.

PLS methods extract latent variables using its standard form which is based on the NIPALS algorithm, taking into account both X and Y. The extracted latent variables are from directions which maximize the empirical covariance between latent or score vectors on the X and the Y block as explained in the PLS algorithm subsection. The extracted components are then deflated by subtracting their rank-one approximations based on the score vectors.

The vectors of loadings are then computed by regressing X on t and Y on u to obtain the coefficients as shown below;

$$p=X^Tt/(t^T t) \qquad \text{and} \qquad q=Y^T u/(u^Tu)$$

Therefore, being an iteration process, the deflation is done after iteration of individual block matrices using the equivalent score and loading vectors. Each iteration of the X and Y matrices are deflated as shown below

$$X=X-tp^T \quad \text{and} \quad Y=Y-uq^T$$

Since the relationship between X and Y is asymmetric, it first assumed that the score vectors or the latent vectors are good predictors of Y and Secondly, it is assumed that the scores vectors t and u relate linearly; that is,

$$U = TD + H \qquad (3)$$

D and H represent the diagonal and residuals respectively. The asymmetric assumption of linearity is then changed into a deflation scheme. In solving linear regression problems, the linearity between the scores vectors assumption in equation 3 is combined with the decomposition matrix Y, equation 2 which is denoted by;

$$Y = TDQ^T + \left(HQ^T + F\ \right)$$

Which formulate the equation below.

$$Y = TC^T + F^* \qquad (4)$$

$C^T=DQ^T$ now represent the matrix of regression coefficients whereas $F^*=HQ^T+F$ is the residual matrix. On decomposing Y using OLSR with orthogonal predictors T we obtain equation 4. Considering the orthonormalised score vectors t i. $T^TT=I$ and $C=Y^TT$ which is not scaled to length one weight vector c, equation 4 is then in term of the original predictors of X. Using the relationship below,

$$T = XW(P^T W)^{-1} \qquad (5)$$

Fixing this relation into equation 4 we get,

$$Y = XB + F^* \qquad (6)$$

Where B is the matrix of regression coefficients

$$B = W(P^T W)^{-1} C^T = X^T U(T^T X X^T U)^{-1} T^T Y$$

3.5 PLS Algorithm

The PLS algorithm starts with scaling, and centering data, X and Y, the it continues as follows,

Let S_x be the variance-covariance of a matrix X denoted by

$$S_X = E[(X - E[X])(X - E[X])^T] \quad (7)$$

If $X \in \mathbb{R}^{n \times k}$ then $S_X \in \mathbb{R}^{k \times k}$

Let $S_{X,Y}$ be the cross-covariance matrix between two matrices $X \in \mathbb{R}^{n \times k}$ and $Y \in \mathbb{R}^{m \times k}$ be defined by

$$S_{X,T} = E[(X - E[X](Y - E[Y])^T] \quad (8)$$

And $S_{X,Y} \in (k \times m)$

The covariance matrix for the regression between the two matrices X and Y is defined as

$$S = \begin{vmatrix} S_X & S_{X,Y} \\ S_{X,Y}^T & S_Y \end{vmatrix}$$

Therefore $s \in \mathbb{R}^{(k+m)\times(k+m)}$ and whose elements are the scalar variances and covariances within and between the columns of X and Y. PLS intention is to maximize the covariance between the regressor matrix and response by finding the vectors to project the cross-covariance S_{XY} between the regressor matrix X and the response matrix Y. In order to make

the cross-covariance function scalar, PLS algorithm starts by using weights vectors to project the regressor and response matrices into scores vectors

$t = Xw^T$ and $U = Yc$

Where here the weights vectors w and c are unit length. The algorithm can therefore be stated as an optimization problem by maximizing

$f(w, c) = t^T U$ 9

With respect to w and c, where

$$t = Xw^T$$

$$u = Yc$$

Subject to

$$W^T W = 1$$

$C^T C = 1$ (10)

Since $t^T u = w^T X^T Yc$the Langangian function is

$L(w, c) = w^T X^T Yc + \lambda_1(1 - w^T w) + \lambda_2(1 - c^T c)$ (11)

The solution to the maximum cross-covariance function problem is the point where all four differential with respect to, w, c, λ_1 and λ_2 are zero as shown below.

$\frac{\partial L}{\partial w} = X^T Yc - 2\lambda_1 w = 0$ (12)

$\frac{\partial L}{\partial c} = Y^T Xw - 2\lambda_2 c$ =0 (13)

$\frac{\partial L}{\partial \lambda_1} = 1 - w^T w$ = 0 (14)

$\frac{\partial L}{\partial \lambda_2} = 1 - c^T c$ = 0 (15)

The partial differentials equations 12 and 13 are vector differecials, when we compare the two equations, they are the same apart from the transpose, and therefore the multipliers are equal $\lambda_1 = \lambda_2$ .This turns the solution into a singular value decomposition form (SVD).

$$cY^T Xw = wX^T Yc = s$$

where s is singular value associated with $Y^T X$ or equivalently $X^T Y$. Since the optimization function from equation 9 to be maximized is $f(w, c) = t^T u = w^T X^T Yc$ the unique solution for s is the maximal singular value of $X^T Y$ and$Y^T X$.W and c vectors that are the solution to the optimization are the left and right vectors of the SVD form. To calculate the weights vectors, the SVD of the cross-covariance matrix is computed,

$$USV^T = SVD(X^T Y)$$

The weights vector w and c are then the first columns of U and V. The other optimization constraints of unit length on both weights vectors w and c are also consistent with this SVD solution as the columns of U and V from a singular value decomposition are all orthonormal by definition. The cross-covariance decomposed for the general form of PLS for multivariate responses as equation 16, is generally not square, so two weights vectors and consequently two scores vectors are required. Deflation is then done by subtracting the variance explained for each latent variable from the X regressor matrix, Y responses or

their cross-covariances. After J latent variable iterations, the fitted values of X and Y, for a general case of PLS multivariate responses are,

$$\hat{X}_J = \sum_{a}^{J} t_a p_a^T$$

$$\hat{Y}_J = \sum_{a=1}^{J} u_a q_a^T$$

where the X and Y loadings p and q are row vectors denoted by,

$$p_a = \frac{t_a^T X_a}{t_a^t t_a}$$

$$q_a = \frac{t_a^T Y_a}{t_a^T t_a}$$

Therefore, the deflation for this Basic PLS is,

$$X_{a+1} = X_a - \hat{X}_a = X_a - t_a p_a^T$$

$$Y_{a+1} = Y_a - \hat{Y}_a = Y_a - t_a q_a^T$$

After deflation, the estimation of the next components starts again using the deflated X and Y matrices. Vectors are saved, as columns in matrices after iterations. The T scores are then used to calculate the regression coefficients, and later converted back to the dominion of the original variables by multiplying with matrix W.

CHAPTER FOUR

DATA ANALYSIS, RESULTS AND DISCUSSION

4.1 Introduction

This chapter highlights the findings of the study and discussion of the results.

4.2 Data analysis software.

PLS-R demand complex computation and therefore, the availability of software determine its application depends. In chemistry, SIMCA-P which was developed originally by Wold, and UNSCRAMBLER are mainly used. SPM, which is one of the most widely used programs in brain imaging field, has recently (2002) integrated a PLS regression module.SAS PROC PLS and R software that are probably the most easily available programs. In this study, data analysis will be done using the R software version 3.1.1(2014-07-10).

4.2.1Correlations of Variables

Mathematics, English and Kiswahili subject, show a higher correlation with Physics, Chemistry and Biology subjects, compared to gender and the school type as indicated in the figure 1.

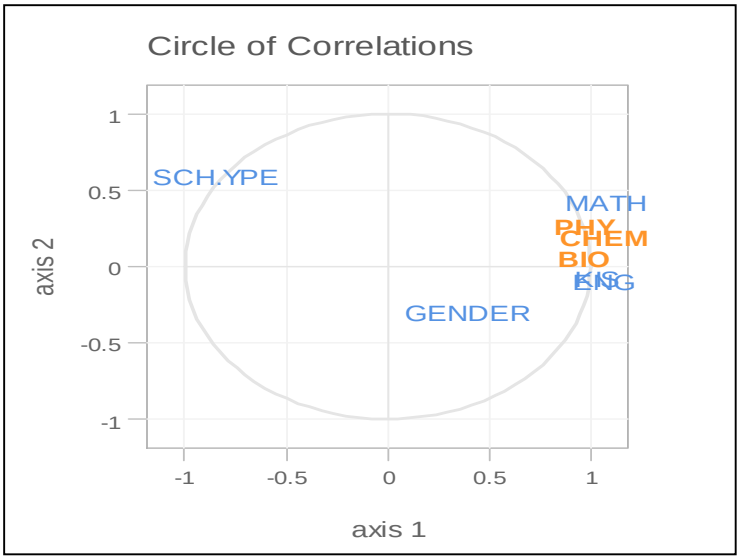

Figure 4. 1: Correlations between physics, biology and chemistry and the latent vectors

4.2.2 The Variance (R^2) of X and Y explained by the LV

The coefficient of determination (R^2) is the statistical measure that is used to determine the percentage of variance explained by the variables.Q^2 is used for cross validation in which it explains how well the variables are valid in predicting new observations.

Table 4.1: Percentage of explained variance of Y and X

Latent vectors	Percentage of Explained Variance of X	Cumulative % of Explained Variance of X	Percentage of Explained Variance of Y	Cumulative % of Explained Variance of Y
1	0.523	0.523	0.674	0.674
2	0.127	0.650	0.036	0.710
3	0.164	0.814	0.003	0.714

Table 4.2: Cumulative variance of cross-validation explained by the latent vectors

Latent vector	Cumulative Q^2 of BIO	Cumulative Q^2 of PHY	Cumulative Q^2 of CHEM	Cumulative Q^2
1	0.682	0.652	0.688	0.674
2	0.686	0.720	0.724	0.710
3	0.691	0.725	0.724	0.710

From table 2, we find out that the first three components or latent variables explain 81% of the variance of X and 71% of Y. These components predict 69% of the biology subject, 73% and 72% of physics and chemistry subjects respectively. Overall, the constructed factor models are of good fit since 71% of the predicted observations are valid.

4.2.3 The standard coefficients of PLS regression.

The examination of the matrix of the coefficient shown in table 4, suggest that gender is negatively correlated to the performance in the science subjects. English and Kiswahili subjects are mainly responsible for the performance in Biology. Although the performance in physics and chemistry is mainly influenced by mathematics, there is a positive correlation with the language subjects which suggest that they also do play a role in the performance of the two subjects.

Table 4.3: The coefficient matrix when 3 latent vectors are used

	Biology	Physics	Chemistry
Gender	-0.096	-0.053	-0.075
School Type	-0.079	0.067	0.0128
English	0.262	0.196	0.225
Kiswahili	0.326	0.198	0.254
Mathematics	0.308	0.574	0.479

4.2.4 Variable of Importance for Projection (VIP)

For assessment of relative variable importance in each PLSR model, the information content of each variable is assessed by its VIP. The three subjects, English, Kiswahili and Mathematics were found out to be of importance in modeling the performance in sciences. Gender and type of school were found to be not important in the model.

	1	**2**	**3**
Gender	0.006	0.167	0.203
School type	0.574	0.642	0.641
English	1.185	1.162	1.159
Kiswahili	1.233	1.204	1.203
Mathematics	1.321	1.327	1.326

Table 4.4: Variable of Importance for Projection.

CHAPTER FIVE

CONCLUSION DISCUSION AND RECOMMENDATION

5.1 Conclusion and Discussion

The analysis shows that there exists a relationship between English, Kiswahili and Mathematics and the science subjects. Literacy which is indicated by the English and Kiswahili contributed more in the achievement of the processing skills of biology at 0.326 and 0.345 compared to the numeracy skills indicated by mathematics at 0. 247.The processing skills of Physics and Chemistry were mainly influenced by the numeracy skills at 0.479 and 0.405 respectively. Although the performance in Physics is mainly influenced by mathematics, languages do play a role since the correlation is high and positive.

Gender is negatively correlated to the achievement of the science process skills. The type of school has a positive correlation with performance in Biology and a positive correlation with the performance of both the physics and chemistry subjects.

5.2 Recommendation

The study have shown that the achievement in languages and mathematics contribute to the achievement in sciences. The study therefore recommends that to address the dismal performance in sciences at KCSE, the government should not only focus on mathematics and sciences but should also focus on the development of literacy skills in English and Kiswahili. Enhancement of literacy skills in English and Kiswahili should one of the activities in the capacity development activity in mathematics and sciences. Policy-makers should consider planning for capacity development training for English and Kiswahili teachers.

REFERENCES

Bjon-Helge Mevik, Ron Wehrens (2007), Principal component and Partial Least Squares Regression in R. *Journal of Statistical Software*. Volume 18, Issue2.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modeling Based on Generalized Linear Models*. Springer Series in Statistics, Berlin.

Gilberto J. Cuevas (1984) *Journal for Research in Mathematics Education, Minorities and Mathematics.*

Joe, F. Hair Jr, Marko Sarstedt, Lucas Hopkins , Volker G. Volume 58,pp.109-139 Kuppelwieser , (2014). "Partial least squares structural equation modeling (PLS-SEM): An emerging tool in business research*", European Business Review, Vol. 26 Issue: 2pp.106 – 121*

Kalivas, J. (1999). Interrelationships of multivariate regression methods using eigenvector basis sets. *Journal of Chemometrics, 13:111-132*

KNEC (2007) KCSE (2006) Examination Report. Nairobi, Government Printer

Lanczos, C. (1950). An iteration method for the solution on the eigen value problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards, 45(4):252{282. Research paper 2133.*

Luis, M. Carrascal, Ismael Galva´n and Oscar Gordo, (2009) *Partial least squares regression as an alternative to current regression methods used in ecology*.ES28006 Madrid, Spain.

Manne, R. (1987). *Analysis of two partial-least-squares algorithms for multivariate calibration*. Chemometrics and Intelligent Laboratory Systems, 2:187{197.

Martens, H. and Naes, T. (1989*). Multivariate Calibration.* John Wiley & Sons, New York.

Meiers, M. (2010). Language in mathematics classroom. The Digest, NSWIT. Retrieved Oct 2015, from http://www.nswteachers.nsw.edu.au.

Svante wold, Michael Sjostron, Lennart Ericksson (2001). PLS-Regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems.

Wold, H. (1975) Path models with latent variables: the NIPALS approach, in: H.M. Blalock (Ed.), *Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building*. New York, Academic Press, 307-335.

Wold, H. (1975). *Soft modelling by latent variables*: the nonlinear iterative partial least squares approach. In Gani, J., editor, Perspectives in Probability and Statistics, Papers in Honour of M.S. Bartlett, pages 520{540, London. Academic Press.

Wong, K.K-K. (2011) Book Review: Handbook of Partial Least Squares: Concepts, Methods and Applications, International Journal of Business Science and Applied Management, Volume6, Issue 2, pp. 53-54.