UNIVERSITY OF NAIROBI

SCHOOL OF MATHEMATICS

Assessment of quality of life of breast cancer patients using principal component analysis

CHEBET, Sharon

I56/60588/2013

June 2014

A research project submitted to the Department of Mathematics in the School of Mathematics in partial fulfillment of the requirements for the award of the degree of Master of Science in Biometry of the University of Nairobi.

Declaration

This thesis is my original work and has not been presented for examination in any other university.

Signature…… …… Date……

CHEBET, Sharon

This project has been submitted for examination with the approval of University supervisor.

Dr. Nelson Owuor

Department of Mathematics Signature…….

University of Nairobi Date…….

Acknowledgements

I wish to acknowledge and appreciate my supervisor, Dr. Owuor for the supervision, constructive comments, encouragement and guidance he provided me throughout my research. His invaluable support and advice is unforgettable.

I am grateful to Higher Education Loans Board (HELB) for the award of scholarship. The financial contribution is highly appreciated.

Of course this work would not have been a success without the kind support, love and prayers of my beloved husband and my family. Thank you.

Finally, and most importantly, I thank the Lord Almighty, for His love and his abundant grace. He is the creator of all things. I forever worship Him.

Abstract

Background: The most commonly diagnosed cancers worldwide are those of the lung, breast and colorectum. Breast cancer is the most common cancer in women worldwide. In 2012, 1.7 million women were diagnosed with breast cancer and there were 6.3 million women alive who had been diagnosed with breast cancer in the previous five years. Since 2008, breast cancer incidence has increased by more than 20%, while mortality has increased by 14%. A cross-sectional study was carried out at the Haemato-oncology and Cancer Treatment Centre of Kenyatta National Hospital to assess the quality of life (QOL) of breast cancer patients. Patients receiving cancer treatment were consecutively recruited at a rate of 20 patients per week until the required sample size of 140 was achieved. Each patient was interviewed using a validated tool for assessing quality of life in cancer patients – European Organization for Research and Treatment of Cancer, Quality of Life Questionnaire (EORTC QLQ-C30). Principal component analysis (PCA) was used to identify a single variable that indicates QOL, before applying logistic regression to assess the predictors of QOL. Nonlinear PCA analysis resulted in total percentage variance accounted for (PVAF) of 46.8%. The median QOL score was 2.45. The mean age of the respondents was 49.4 ±10.2 years. 61.8% of the respondents were in their late stages of the disease. Thirty eight percent respondents were on chemotherapy, 27.5% on radiotherapy, 20.4% on tamoxifen and 14.1% were on surgery. The study found out that surgery (p=0.010) and tamoxifen treatments (p=0.001) are statistically significant predictors. Ages, marital status, education, parity, stage of the disease and radiotherapy treatment are not significant predictors. The study concludes that breast cancer patients have poor quality of life, and those in the late stage of the disease are more likely to have poor quality of life compared to those in the early stage. Patients receiving surgery and taxomifen reported lower QOL score. This study will therefore be used in management of QOL of breast cancer patients by directing innovative interventions that improve quality of life of patients.

Key words: Breast cancer, PCA, QLQ-C30, EORTC, PVAF

Table of Contents

List of tables

List of Figures

Abbreviations

ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
CATPCA	Categorical principal components analysis
CI	Confidence interval
EORTC	European Organization for Research and Treatment of Cancer
PCA	Principal Component analysis
Pr	Probability
PVAF	Percentage variance accounted for
QLQ-C30	Quality of life questionnaire
QOL	Quality of Life
VAF	Variance accounted for
WHO	World Health organization

1. INTRODUCTION

This chapter presents; the background of the study, statement of the problem, objectives and significance of the study.

1.1. Background of the study

Cancer is an abnormal growth of cells which tend to proliferate in an uncontrolled way and, in some cases, to metastasize (spread). Cancer is not one disease but a group of more than 100 different and distinctive diseases. Cancer can involve any tissue of the body and have many different forms in each body area. Most cancers are named for the type of cell or organ in which they start [1].

The most commonly diagnosed cancers worldwide are those of the lung (1.8 million, 13.0% of the total), breast (1.7 million, 11.9%), and colorectum (1.4 million, 9.7%). The most common causes of cancer death are cancers of the lung (1.6 million, 19.4% of the total), liver (0.8 million, 9.1%), and stomach (0.7 million, 8.8%). According to [2], an estimated 14.1 million new cancer cases and 8.2 million cancer - related deaths occurred in 2012, compared with 12.7 million and 7.6 million, respectively, in 2008. Prevalence estimates for 2012 show that there were 32.6 million people (over the age of 15 years) alive who had had a cancer diagnosed in the previous five year**.** Projections based on the [2] estimates predict a substantive increase to 19.3 million new cancer cases per year by 2025, due to growth and ageing of the global population. More than half of all cancers (56.8%) and cancer deaths (64.9%) in 2012 occurred in less developed regions of the world, and these proportions will increase further by 2025 [2].

Breast cancer is the most common cancer in women worldwide (it represents one in four of all cancers in women [2]), with male to female ratio of 1:100. Rates of breast cancer around the world vary. The highest incidence of breast cancer was in Northern America and Oceania; and the lowest incidence in Asia and Africa [3]. These low rates have been attributed to low screening rates and incomplete reporting. Other reasons are; rapid societal and economic changes, the shift towards lifestyles typical of industrialized countries leads to a rising burden of cancers associated with reproductive, dietary, and hormonal risk factors. In 2012, 1.7 million women were diagnosed with breast cancer and there were 6.3 million women alive who had been diagnosed with breast cancer in the previous five years. Since the 2008

estimates, breast cancer incidence has increased by more than 20%, while mortality has increased by 14% [2].

The government of Kenya has two cancer registries; The Nairobi Cancer Registry located at the Kenya Medical Research Institute (KEMRI) that captures data from Nairobi and its environs and the Eldoret Cancer Registry located at the Moi Teaching and Referral Hospital which serves mainly the North Rift and Western provinces of Kenya.

Nairobi Cancer Registry is a population-based registry that was established in the year 2001 at KEMRI, Nairobi. Population-based registry records all new cases in a defined population with emphasis on epidemiological research, and evaluation of health services for prevention, diagnosis and treatment of the disease. The newly released report indicated that the top three common cancer of the male is prostate cancer, oesophagus cancer and stomach cancer. For females, cancer of the breast, cancer of the uterine cervix, and stomach and oesophageal cancer were reported as the most common. Generally, cancers of the breast and prostate cancer contribute to the highest burden of cancer in the country.

The Eldoret Cancer Registry is located at the Moi Teaching and Referral Hospital which serves mainly the North Rift and Western provinces of Kenya. The registry was established in 1999. By January 2012, the topmost cancer of men was oesophagus cancer, skin cancer, non-Hodgkin lymphoma cancer and prostate cancer while cancer of cervix, breast cancer and oesophagus and skin cancer were the most common. The ratio of male to female is approximately one to one. Currently, there is no national cancer registry in Kenya.

WHO [4] definition of quality of life is individuals' perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns. It is a broad ranging concept affected in a complex way by the person's physical health, psychological state, level of independence, social relationships, personal beliefs and their relationship to salient features of their environment. Quality of life assessment can be used in diagnosis, predicting prognosis, patient monitoring, clinical decision-making, communication, and treatment. It helps in analysis of quality of healthcare and in identifying areas to improve.

Several instruments measuring quality of life of cancer patients have been developed. One of them is European Organization for Research and Treatment of Cancer quality of life questionnaire (EORTC QLQ-C30), the first version (QLQ-C36) was developed in 1987 and

the current version QLQ-C30 was developed in 2000. The QLQ-C30 is composed of both multi-item scales and single-item measures. These include five functional scales, three symptom scales, a global health status / QOL scale, and six single items. Each of the multi-item scales includes a different set of items - no item occurs in more than one scale [5]. Other instruments are WHO Quality of Life-BREF (WHOQOL-BREF) initiated in 1991 [6], Spitzer Quality of Life Index (QLI) (1981), Rotterdam Symptom Check List (RSCL)-(1990), Functional Living Index-Cancer (FLIC) - 1984, Functional Assessment of Cancer-1993 and Therapy-General (FACT-G) [7] (citing [8], [9], [10]).

1.2. Statement of the problem

Incidences of breast cancer have been increasing in most regions of the world. Incidence rates remain highest in more developed regions, but mortality is relatively much higher in less developed countries due to a lack of early detection and access to treatment facilities. Breast cancer is the most common cancer in women. In 2012, 1.7 million women were diagnosed with breast cancer and there were 6.3 million women alive who had been diagnosed with breast cancer in the previous five years. Since the 2008 estimates, breast cancer incidence has increased by more than 20%, while mortality has increased by 14%. National Cancer Control Strategy (2011 – 2016) has been created in Kenya and one of its strategies is treatment of cancer. The goal of this strategy is to ensure the best possible quality of life for cancer patients. There is therefore an urgent need to develop effective and affordable approaches to the early detection, diagnosis, and treatment of breast cancer and assessing the patients' quality of life can be used for that purpose. One important benefit of quality of life (QOL) assessment is to encourage shared decision making and to facilitate communication between physicians and patients by providing feedback to the patients regarding their progress, goals, and expectations. QOL measures can also be used for monitoring disease progression (i.e., survival) or response to treatment (i.e., toxicity, side effects, and adverse effects). Understanding the effect of breast cancer treatment on a patient's QOL has been a central clinical and research question [11]. This study on quality of life assessment of breast cancer patients is therefore important because it will be used as a preventive intervention and inform clinicians about the patient's illness as well as how certain treatments may affect the QOL of that patient.

1.3. Objectives

The overall objective is to identify a single variable that indicates the health related quality of life of breast cancer patients at Kenyatta National Hospital.

The specific objectives are:

1. To determine whether stage at diagnosis and type of treatment received influences the quality of life of breast cancer patients
2. To determine whether socio-demographic characteristics influence quality of life of breast cancer patients

1.4. Significance of the study

Breast cancer is the most common cancer in women. Breast cancer incidence and mortality are on the increase. Many researchers have shown that younger women are mostly affected by breast cancer [12][13][14]. This has an adverse effect on the country's economy because young women are mostly involved in activities that boost the economy.

This study will provide information concerning the treatment outcome that will be used for policy formulation by the Ministry of Health. This study will also bring out clearly the information concerning the association on stage of the disease and the quality of life of the patients which will be used as a mechanism for the promotion of early detection. Finally this study will contribute to the knowledge based with regard to socio-demographics factors that affect the quality of life of breast cancer patients.

Chapter two presents the literature review and the summary of the literature review. Chapter three discusses the methodology; the study sample, the measure and description of statistical methods used (principal component analysis and logistic regression). Chapter four presents analysis and results and chapter five presents the conclusions and recommendation.

2. LITERATURE REVIEW

Various studies have been carried out to determine factors affecting the quality of life of breast cancer patients. These studies are described below.

A study in Curitiba Brazil [12] on two medical centres revealed that predictive factors of fatigue included younger age (p=0.024), presence of pain (p=0.000), dyspnea (p=0.006), insomnia (p=0.015) and nausea and vomiting (p=0.036). Socio-economic characteristics (ethnicity, educational level, marital status, number of children, employment status and individual income per month) were not related to fatigue. The statistical method used was logistic regression.

A study was done in Kuwait in western Asia [13] to examine the association of health-related quality of life with socio-demographic characteristics, stage of disease, type of treatment received in the past, and duration since last treatment. The statistical methods employed were by Pearson's correlations, t-tests and one-way Analysis of Variance (ANOVA) and multiple (stepwise) regression analyses. The study revealed that factors associated with health-related quality of life were age, stage of cancer (the significant associations were only for three scales; role functioning, diarrhoea and future perspective), radiotherapy and fatigue. The functional scale scores were more important in predicting functional scales, than the scores of the symptom scales, while social functioning accounted for the highest proportion of variance in global quality of life**.** The study also found out that physical and role functioning are highly correlated.

A study was done in India [14] to assess the determinants of quality of life of breast cancer patients. Data was analysed using ANOVA and multinomial logistic regression. Younger women (<45 years), women having unmarried children, nodal and/or metastatic disease, and those currently undergoing active treatment showed significantly poorer quality of life scores in the univariate analysis. However logistic regression analysis indicated that religion, stage, pain, spouse education, nodal status, and distance travelled to reach the treatment centre as indicative of patient quality of life**.**

A study to assess the variables associated with quality of life of Sudanese women cancer outpatients was done using correlations, analysis of covariance (ANCOVA), t-test and one way ANOVA [15]. The study revealed that the significant covariates were marital status, occupation and education of the patient (being married, educated, and formally employed had

significantly higher quality of life). The duration of illness was also significantly correlated with all the patient's quality of life domain scores except for the spiritual domain (longer duration of illness had higher quality of life). Patients on chemotherapy were not associated with quality of life while patients on radiotherapy were significantly correlated with physical health, psychological health, social relations and spiritual domain. Using multiple (step-wise) regression analysis, duration of illness and patient feeling currently ill were significant predictors of patient's quality of life**.**

A study done in Iraq [16] to evaluate the impact of adjuvant therapy on quality of life in patients with breast cancer, and to find out the differences in the quality of life between patients receiving chemotherapy and radiation therapy revealed that quality of life of patients receiving chemotherapy and radiotherapy were impaired but there were no significant differences between the two groups regarding the psychosocial wellbeing domain. There were significant differences between them for the physical complaints (radiotherapy patients had low mean score while chemotherapy patients had better score) and for the daily activities domain (chemotherapy patients had low mean score and radiotherapy had medium mean score). T-test was used to compare mean between the two groups.

A study was done in the two large metropolitan cities (Los Angeles - California, and Washington DC) in USA [17] to describe the occurrence of fatigue in a large sample of breast cancer survivors relative to general population norms and to identify demographic, medical, and psychosocial characteristics of fatigued survivors. Pearson correlation, chi square, t-test, ANOVA and logistic regression were used. The bivariate analysis showed that women in the fatigued group were slightly younger, had a lower yearly income, and were less likely to be married or in a significant relationship than those in the non-fatigued group. Ethnicity, educational level, and employment status were not associated with fatigue. There was a modest association between fatigued women and type of treatment received but no association with taxomifen group. The results from logistic regression showed that depression and pain were the strongest predictors. Type of treatment received was not a significant predictor.

2.1. Summary of literature review

The summarized studies revealed that age, pain, dyspnea, insomnia, nausea and vomiting, religion, spouse education, nodal status, distance travelled to reach treatment centre, duration of illness, patients feeling currently ill, depression and stage of the disease are significant

predictors of quality of life. The studies further found that ethnicity, education, marital status, parity, employment status, income and treatment received are not significant predictors. These studies were done in India, Iran, USA, Asia, Brazil and Sudan. To the best of the knowledge of the researcher no studies have been done in Kenya.

Statistical methods that were most commonly used are t-test, ANOVA, chi square, correlation and logistic regression. Most studies used some of the health related quality of life as dependent variables while others used global health status / quality of life scale. To the best of the knowledge of the researcher no studies have been done using principal component analysis.

3. METHODOLOGY

This chapter present the study sample, the measure and description of statistical methods used.

3.1. Study sample and recruitment

A cross-sectional study between June and September 2011 was carried out at the Haemato-oncology and Cancer Treatment Centre of Kenyatta National Hospital. Records of two hundred breast cancer patients receiving cancer treatment were consecutively sampled. Fifty patients did not meet the inclusion criteria – forty patients were not on any treatment modality while ten patients did not have any documented breast cancer histology in their files. Out of those who met the inclusion criteria, eight patients declined consent, leaving one hundred and forty two (142) patients who were then recruited. Each patient was interviewed using a validated tool for assessing quality of life in cancer patients – the EORTC QLQ-C30. Patients were eligible for recruitment if they meet the following inclusion criteria: age above 18 years, written informal consent, and diagnosis of breast cancer by tissue histology or cytology and those who are on-going or completed any standard modality of breast cancer treatment within the preceding 6 months.

3.2. Measure

EORTC QLQ-C30 is a tool for assessing quality of life of cancer patients. The QLQ-C30 is composed of both multi-item scales and single-item measures. Multi-item scales are five functional scales (physical, role, cognitive, emotional and social functioning), three symptoms scale (pain, fatigue and nausea & vomiting and global health status / quality of life scale. The single-item measures consist of Dyspnoea, Insomnia, Appetite loss, Constipation, Diarrhoea and Financial difficulties. Each of the multi-item scales includes a different set of items. All the items except global health status / quality of life scale are 4 point scale (1=Not at All, 2= A little, 3= Quite a Bit, 4= Very Much). Global health status / quality of life scale is a general measure of quality of life and it consists of 7 point scale. A higher score for functional scales represents high or healthy level of functioning, a high score for the global health status / QOL represents a high QOL and a high score for symptoms scale / items represents a high level of symptomatology / problems.

3.3. Description of the statistical methods

This section discusses the description of both linear and nonlinear principal component analysis and logistic regression.

3.3.1. Principal component analysis

3.3.1.1. Introduction

Principal component analysis is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set. It describes the variation in a set of correlated variables $(x_1, x_2, \ldots . x_q)$ in terms of a new set of uncorrelated variables $(y_1, y_2, \ldots . y_p)$, each of which is a linear combination of the x variables.

3.3.1.2. Geometric basis of the principal component analysis

Geometrically, the objective of PCA is used to identify new sets of orthogonal axes such that

- The coordinates of the observation $(x_1, x_2, \ldots . x_q)$ with respect to each of the axes give the values of the new variables $(y_1, y_2, \ldots . y_p)$
- Each new variable is a linear combination of the original variables
- The first new variable accounts for the maximum variance in the data, the second new variable accounts for the maximum variance that has not been accounted for by the first variable and the p^{th} new variable accounts for the variance that has not been accounted for by the p-1 variables
- The p new variables are uncorrelated

PCA deals with a single sample of n observation vectors $(x_1, x_2, \ldots . x_n)$ that form a swarm of points in a p-dimensional space (Figure 1). Since the variables $(x_1, x_2, \ldots . x_q)$ are correlated, the ellipsoid swarm of points is not oriented parallel to any of the axes represented by $(x_1, x_2, \ldots . x_q)$. We therefore find the natural axes of the swarm of points with origin at $\bar{x}$, the mean vector of $(x_1, x_2, \ldots . x_q)$. This is done by translating the origin to $\bar{x}$ and then rotating the axes. After rotation so that the axes become the natural axes of the ellipsoid, the new variables (principal components) will be uncorrelated. The axes can be rotated by multiplying each x_i by an orthogonal matrix **A,** $y_i = \boldsymbol{A}x_i$

Since **A** is orthogonal, $\boldsymbol{A'A} = 1$, and the distance to the origin is unchanged, then;

$$y_i{}'y_i = (\boldsymbol{A}x_i)'(\boldsymbol{A}x_i) = x_i{}'\boldsymbol{A'A}x_i = x_i{}'\boldsymbol{I}x_i = x_i{}'x_i$$

Thus an orthogonal matrix **A** transforms x_i to a point y_i that is the same distance from the origin, and the axes are effectively rotated. Finding the natural axes of the swarm of points is equivalent to finding the orthogonal matrix **A**. The covariance matrix of y_i (S_y) is a diagonal matrix, such that $S_y = \boldsymbol{ASA'}$, where $\boldsymbol{S}$ is the sample covariance matrix of $(x_1, x_2, \ldots . x_q)$ and **A** is the transpose of the matrix whose columns are normalized eigenvectors of $\boldsymbol{S}$.

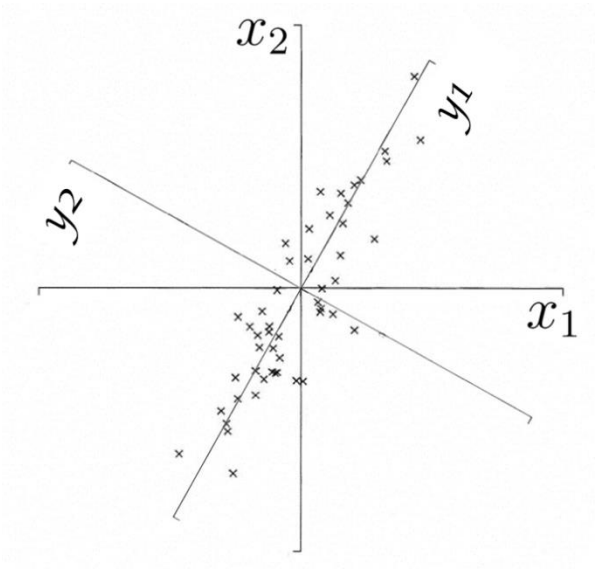

Figure 1. Geometric representation of principal component analysis

3.3.1.3. Algebraic basis of the principal component analysis

The p-dimension principal components may be defined by the following equations;

$$y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1q}\,x_q$$

$$y_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2q}\,x_q$$

.

.

$$y_p = a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pq}\,x_q \qquad (3.2.1)$$

Where:

$\boldsymbol{y_1}$ represents the first principal component which is the linear combination of x-variables that has maximum variance (among all linear combinations), so it accounts for as much

variation in the data as possible. It is subject to $\underline{a}'_1\underline{a}_1 = 1$, ($\underline{a}'_1$ = transpose of $\underline{a}_1$), The restriction is placed on the coefficients because the variance of y_1 could be increased without limits by simply adding the coefficients $\underline{a}_1$.

$\boldsymbol{y_2}$ represents the second principal component which is a linear combination of x-variables that accounts for as much of the remaining variation as possible subject to the following two conditions; $\underline{a}'_2\underline{a}_2 = 1$ and $\underline{a}'_2\underline{a}_1 = 0$. $\underline{a}'_2\underline{a}_1 = 0$ ensures that y_1 and y_2 are uncorrelated.

In general, all subsequent principal components have this same property – they are linear combinations that account for as much of the remaining variation as possible and they are not correlated with the other principal components.

$\boldsymbol{y_j}$ represents the j[th] principal component which has the greatest variance subject to the following conditions: $\underline{a}'_j\underline{a}_j = 1$ and $\underline{a}'_j\underline{a}_i = 0 \ ; \ i < j \in p$.

$\boldsymbol{\underline{a}_j}$ are called the component loadings. The elements of $\underline{a}_1, \underline{a}_2, \ldots . \underline{a}_q$ correspond to the eigenvector (**x**) of ***S*** for the respective eigenvalues $\lambda_1,\lambda_2,\ldots,\lambda_q$.

$\boldsymbol{\underline{x}_i}$ are the observed variables.

In matrix notation;

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_p \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \ldots . & a_{1q} \\ a_{21} & a_{22} \ldots . & a_{2q} \\ . & . & . \\ . & . & . \\ a_{p1} & a_{p2} \ldots . & a_{pq} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_q \end{bmatrix}$$

or;

$$y_1 = \underline{a}'_1\underline{x}$$

$$y_2 = \underline{a}'_2\underline{x}$$

.

$$y_p = \underline{a}'_p\underline{x} \qquad (3.2.2)$$

Following the PCA weights estimations conditions i.e. $\underline{a}'_j\underline{a}_j = 1$ and $\underline{a}'_j\underline{a}_i = 0 \ ; \ i < j \in p$, the weights can be obtained by the eigenstructure of the covariance matrix.

3.3.1.4. Eigenvalues and eigenvectors

For every square matrix A, a scalar λ and a nonzero vector **x** can be found such that

$Ax = \lambda x$ where λ is an eigenvalue of A and **x** is an eigenvector of A corresponding to λ

Eigenvectors are those special vectors that are in the same direction as Ax. Almost all vectors change direction when they are multiplied by A. To find λ and **x**, we solve the following characteristic equation: $(A - \lambda I)x = 0$.

If A is $m \times m$ matrix, we will have m eigenvalues such that $\lambda_1 \leq \lambda_2 \leq \ldots, \leq \lambda_m$. The accompanying eigenvectors $(x_1, x_2, \ldots, x_m)$ can be found by substituting eigenvalues into the characteristic equation.

Illustration

Consider the matrix $\boldsymbol{A} = \begin{bmatrix} 3 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 2 & 0 \end{bmatrix}$ and find the eigenvalues and eigenvectors.

$$|A - \lambda I| = 0 \Rightarrow \begin{vmatrix} 3-\lambda & 1 & 1 \\ 1 & 0-\lambda & 2 \\ 1 & 2 & 0-\lambda \end{vmatrix}$$

$$= (3-\lambda)[(0-\lambda)(0-\lambda) - (2)(2)] - (1)[(1)((0-\lambda) - (2)(1)] + (1)[(1)(2) - (0-\lambda)(1)] = 0$$

The resulting eigenvalues are; $\lambda_1 = 4$, $\lambda_2 = 1$ and $\lambda_3 = -2$. To find the eigenvector corresponding to $\lambda_1 = 4$ we solve;

$$\begin{bmatrix} 3-4 & 1 & 1 \\ 1 & 0-4 & 2 \\ 1 & 2 & 0-0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

The eigenvectors corresponding to λ_2 and λ_3 are solved as λ_1. The normalized eigenvectors are;

$$x_1 = \begin{bmatrix} 0.82 \\ 0.41 \\ 0.41 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 0.58 \\ -0.58 \\ -0.58 \end{bmatrix} \text{ and } x_3 = \begin{bmatrix} 0.00 \\ -0.71 \\ 0.71 \end{bmatrix}$$

Spectral decomposition

The eigenvectors of $m \times m$ symmetric matrix **S** are mutually orthogonal. If the m eigenvectors of **S** are normalized and inserted as a column of a matrix **C**, then **C** is orthogonal. The expression $\boldsymbol{S = CDC'}$ for a symmetric matrix **S** is known as the spectral decomposition of **S.**

Where $\boldsymbol{D} = \begin{bmatrix} \lambda_1 & 0 \ldots . & 0 \\ 0 & \lambda_2 \ldots . & 0 \\ . & . & . \\ . & . & . \\ 0 & 0 \ldots . & \lambda_m \end{bmatrix}$ and $\boldsymbol{C}$ is the matrix of the normalized eigenvectors

Since **C** is orthogonal and $\boldsymbol{C'C = CC' = I;} \quad \boldsymbol{C'SC = D}$ is true.

Thus a symmetric matrix **S** can be diagonalized by an orthogonal matrix containing normalized eigenvectors of **A**, and the resulting diagonal matrix contains eigenvalues of **A.**

Orthogonal vectors and matrices

Geometrically, orthogonal vectors are perpendicular to each other. Two vectors **s** and **t** of the same size are said to be orthogonal if $\boldsymbol{s't} = 0$. If $\boldsymbol{s's} = 1$, the vector **s** is said to be normalized. The vector **s** can be normalized by dividing by its length, $\boldsymbol{\sqrt{s's}}$**.**

A matric **C** whose columns are normalized and mutually orthogonal is called an orthogonal matrix. Since the elements of $\boldsymbol{C'C}$ are the products of columns of C, we have

$\boldsymbol{C'C} = 1 \;\Rightarrow\; \boldsymbol{C\,C'} = 1$

It is clear from equation above that $\boldsymbol{C^{-1}} = \boldsymbol{C'}$ for an orthogonal matrix.

Illustration

Consider a 3×3 matrix A and create an orthogonal matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -2 & 0 \end{bmatrix}$$

We first normalized the columns by dividing by their respective lengths, $\sqrt{3}$, $\sqrt{6}$, and $\sqrt{2}$, to obtain

$$\boldsymbol{C} = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{6} & 1/\sqrt{2} \\ 1/\sqrt{3} & 1/\sqrt{6} & -1/\sqrt{2} \\ 1/\sqrt{3} & -2/\sqrt{6} & 0 \end{bmatrix}$$

3.3.1.5. Principal component analysis from correlations and covariance matrix

By definition, the covariance matrix, **S** is given by;

$$\boldsymbol{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1q} \\ s_{21} & s_{11} & \cdots & s_{2q} \\ \vdots & \vdots & \ldots & \vdots \\ s_{q1} & s_{2q} & \cdots & s_{qq} \end{bmatrix}$$ Where s_{11} is the variance of variable x_1 and s_{1q} is the covariance of variable x_1 and x_q.

The correlation matrix, **R** is given by;

$$\boldsymbol{R} = \begin{bmatrix} 1 & \gamma_{12} & \cdots & \gamma_{1q} \\ \gamma_{21} & 1 & \cdots & \gamma_{2q} \\ \vdots & \vdots & \ldots & \vdots \\ \gamma_{q1} & \gamma_{2q} & \cdots & 1 \end{bmatrix}$$ Where γ_{1q} is the correlation coefficient of variable x_1 and x_q

The variance, covariance and correlation can be obtained by the following formula;

- Variance: $s_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Covariance: $s_{xy} = E[(x_i - \bar{x})(y_i - \bar{y})] = \frac{1}{n-1}\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]$
- Correlation: $\gamma_{xy} = \frac{s_{xy}}{\sqrt{s_x s_y}}$ where $s_y = \sqrt{E\ (y_i - \bar{y})^2}$

Principal components analysis is performed on either **S** or **R.** If the variances differ widely or if the measurement units are not the same, the components of **S** will be dominated by the variables with large variances. The other variables will contribute very little. When the variables are standardized, any changes of scale on the variables would not affect the components because **R** itself is scale invariant.

3.3.1.5.1. Variance of the principal components

If the eigenvalues of $\boldsymbol{S}$ are $(\lambda_1, \lambda_2, \ldots, \lambda_q)$ then since $\underline{a}'_j\underline{a}_i = 1$, the variance of the j[th] principal component is λ_j. The total variance of the $\boldsymbol{q}$ principal components will equal the total variance of the original variable i.e. $\sum_{j=1}^{q}\lambda_1 = s^2{}_1 + s^2{}_2 + \ldots + s^2{}_q$.

Consequently, the j[th] principal component accounts for a proportion p_j of the total variation in the original data, where $p_j = \frac{\lambda_j}{\sum_{i=1}^{q} \lambda_1}$. For **R,** $\sum_{i=1}^{q} \lambda_1 =$ total number of variables being analysed.

And the proportion explained by the first **m** principle components ($\boldsymbol{p^m}$) is given by;

$$p^m = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{q} \lambda_1}$$

3.3.1.5.2. Correlations and covariance of variables and principal components

The j[th] principal component is given by; $y_j = a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{ji}x_i \ldots + a_{jq}x_q$. The covariance of the x_i and y_j is therefore given by;

$$cov\,(x_i, y_j) = cov\left(a_{ji}x_i\right) = \lambda_i a_{ji}.$$

And the correlation of x_i and y_j:

$$r_{x_i y_j} = \frac{\lambda_i a_{ji}}{\sqrt{var(x_i)var(y_j)}} = \frac{a_{ji}\sqrt{\lambda_i}}{s_i}$$

However, if the elements are extracted from **R** rather than **S** then; $\gamma_{x_i y_j} = a_{ji}\sqrt{\lambda_i}$ since $s_i = 1$.

3.3.1.6. Steps involved in principal component analysis

Principal component analysis involves the following steps;

- Assessing the suitability of the data
- Choosing the number of components
- Factor rotation and interpretation
- Creating Component Scores or Component-Based Scores

3.3.1.6.1. Assessing the suitability of the data

Before principal component analysis is performed, the following assumptions must be checked.

- **Sample size**

The sample size should be large enough to yield reliable estimates of the correlations among the variables. Various studies have suggested that a sample size of more than 150, [18] (citing [19]) suggested 300 cases.

- **Linearity**

Since principal component analysis is based on correlation, it is assumed that the relationship between two variables should be linear. The graphical method that can be used is the examination of scatterplots, often with a trend line. The variables are linear if scatterplots display a linear shape. The statistical method used is the examination of the significance tests for the pearson correlation coefficients. If the correlation coefficient between two variables is statistically significant, we conclude that the relationship is linear.

Since our data is ordinal, linearity of variables is violated and we shall use nonlinear PCA. It is the nonlinear equivalent of PCA and it deals with variables at their appropriate measurement level, for example, it can treat Likert-type scales at ordinal level instead of numerically. Nonlinear PCA aims at the same goals of linear PCA, and so all mathematical properties of PCA still hold [20]**.**

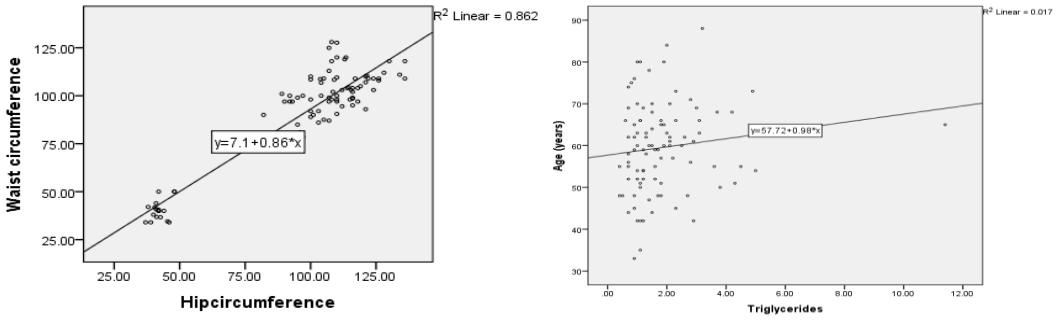

Figure 2: Scatterplot with linear and nonlinear shape

- **Bivariate normal distribution**

Normal is used to describe a symmetrical, bell shaped curve, which has the greatest frequency of scores in the middle, with smaller frequencies towards the extremes. Mean, median and mode are the same and are in the middle of the curve.

Each pair of observed variables should display a bivariate normal distribution. Box plot is used to test for such normality and the shape of the scatterplot obtained should be concentric ellipses.

If two variables (x_1, x_2) are bivariate normal, then individual variables $x_1 \; and \; x_2$ are also normal [20]. There are various graphical and statistical methods used for evaluation of univariate normal data. Graphical methods include; histogram (Figure 3), Q-Q plots (normal data should follow a straight line), Boxplot and stem-and-leaf. The statistical methods are; Kolmogorov-Smirnov statistics and Shapiro Wilk test. Both statistical methods test the null hypothesis that the sample data are not from normal population, a non- significant result represent normality.

Normally distributed variables make the solution stronger but it is not a critical assumption because PCA is an exploratory analysis technique.

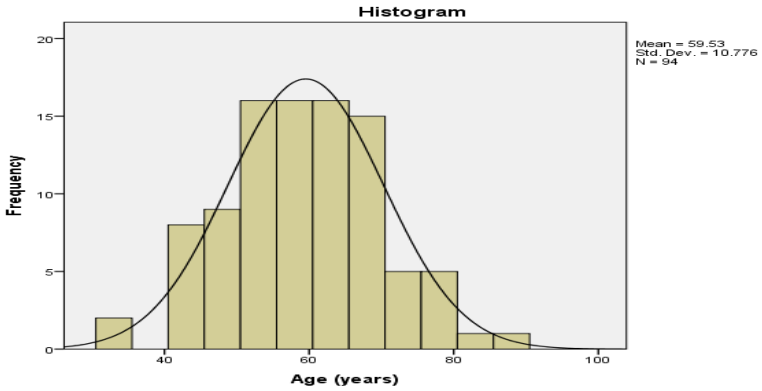

Figure 3: Histogram with normal curve

- **Absence of outliers**

Outliers are cases with values well above or well below the majority of the cases. Principal component analysis is sensitive to outliers. Data needs to be screened and checked for outliers. Histogram can be used to check for outliers, we look at data that are sitting on their own, out on the extremes. Outliers can be removed from the data or recode to a less extreme value.

- **Strength of the inter-correlations among the variables**

PCA requires some degree of collinearity among the variables but not an extreme degree or singularity among the variables. Correlation coefficient (**r**) is the degree of association

between the two variables. Correlation coefficient of +1 or -1 indicates a perfect relationship between two variables while $r = 0$ indicates absence of relationship. Correlation matrix should be inspected for evidence of some coefficient >0.3. If few correlations above 0.3 are found then principal component analysis may not be suitable.

3.3.1.6.2. Choosing the number of components

The number of components extracted is equal to the number of variables being analysed. A decision must be made on how many principal components should be retained in order to effectively summarize the data. The following guidelines have been proposed;

- **Percentage of variance accounted for (PVAF):** Retain just enough components to explain some specified large percentage of total variations of the original variables. Cumulated PVAR between 70% and 90% are usually used.
- **Kaiser's criteria:** Exclude those principal components whose eigenvalues are greater than the average, $\sum_{i=1}^{q} \lambda_i / q$. For a correlation matrix, this average is 1. An eigenvalue represents the amount of variance that is accounted for by a given component. The rationale for this criterion is; each observed variable contributes one unit of variance to the total variance in the data set. Any component that displays an eigenvalue greater than 1.00 is accounting for a greater amount of variance than had been contributed by one variable. Such a component is therefore accounting for a meaningful amount of variance, and is worthy of being retained. On the other hand, a component with an eigenvalue less than 1.00 is accounting for less variance than had been contributed by one variable. The application of this criterion can lead to retaining a certain number of components when the actual difference in the eigenvalues of successive components is minimal. For example, if component 2 displays an eigenvalue of 1.001 and component 3 displays an eigenvalue of 0.999, then component 2 will be retained but component 3 will not; this creates a misleading impression that the third component is meaningless when, in fact, it accounted for almost exactly the same amount of variance as the second component.
- **Scree test:** This involves plotting each of the eigenvalues of the component and inspecting the plot to find a point at which the shape of the curve changes direction and becomes horizontal (Figure 4). It is recommended to retain all factors above the elbow or break in the plot because these factors contribute to most of the explanation of the variance in the data set. Sometimes a scree plot will display several large breaks. When this is the case, one should look for the last big break before the eigenvalues begin to

level off. Only the components that appear before this last large break should be retained. However, this criterion has its own weaknesses as well, very often, it is difficult to determine exactly where in the scree plot a break exists, or even if a break exists at all. When encountered with such a weakness, the use of the scree plot must be supplemented with additional criteria, such as the percentage of variance accounted for criterion, Kaiser's criteria and the interpretability criterion.

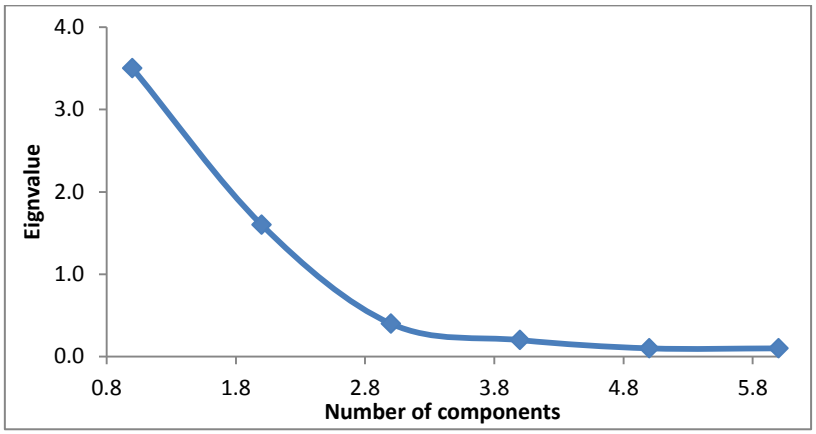

Figure 4: Scree plot for the hypothetical data

- **The interpretability criteria:** These criteria should be supplemented with other criteria. This is done by checking the number of variables and size of loadings on each component. A component with fewer than three items is considered weak and unstable. A crossloading item is an item that loads high on two or more components. If there are few crossloading items, they should be dropped especially when there are several other items with strong / adequate loadings.

3.3.1.6.3. Component rotation and interpretation

The principal components are initially obtained by rotating axes in order to line up with the natural extensions of the systems, where upon the new variables become uncorrelated and reach the direction of maximum variance.

If the resulting components do not have a satisfactory interpretation, they can be further rotated, seeking dimensions in which many of the coefficients of the linear combinations are near zero to simplify interpretation. Rotation is performed when more than one component has been retained in an analysis. Care should be taken in choosing the rotation method so as to retain the property of lack of correlations of the principal components.

There are two main approaches to rotation, resulting in either orthogonal (uncorrelated) and oblique (correlated) component solutions. Orthogonal rotations results in solutions that are easier to interpret and to report, however they require independence (not correlated) of the underlying components. Oblique approaches allow for the factors to be correlated, however they are more difficult to interpret [19]**.** The two approaches result in very similar solutions [21] (citing [22])**.** Within the two approaches there are a number of rotational techniques provided by SPSS. **Orthogonal**; Varimax, Quartimax and Equamax, and **Oblique**; Direct Oblimin and Promax.

Graphical approach of orthogonal rotation

If there are only two components, *m=2,* graphical rotation can be used based on the visual inspection of a plot of component loadings. We choose an angle ∅ through which the axes can be rotated to move them closer to groupings of points. Consider the pairs of component loadings (a_{i1}, a_{i2}), $i = 1, 2, \ldots\ldots, q$, which will be rotated. The new rotated loadings (a^*_{i1}, a^*_{i2}) can be measured directly on the graph as coordinates of the axes or calculated from $a^* = a\boldsymbol{T}$ using

$$\boldsymbol{T} = \begin{bmatrix} cos\ \emptyset & -sin\ \emptyset \\ sin\ \emptyset & cos\ \emptyset \end{bmatrix}$$

Component interpretation

Interpreting the component solution means determining what is measured by each of the retained component; identifying the variable with high loadings and determining what the variables have in common and hence assigning a name to each component. The first decision to be made at this stage is to decide how large a component loading must be to be considered "large." [21] (citing [22]) cite 0.32 as a good rule of thumb for the minimum loading of an item.

3.3.1.6.4. Creating Component Scores or Component-Based Scores

If we have decided we need *m* principal components, then we will calculate the scores on each of these components for each variable in our sample. Having used **S,** the *m* component scores for original $q \times 1$ vector of variable is given by;

$$y_1 = \underline{a}'_1 \underline{x}$$

$y_2 = \underline{a'}_2\underline{x}$

.

.

$y_m = \underline{a'}_m\underline{x}$

If the components are derived from **R,** then $\underline{x}$ should be standardized to have $\mu = 0$ and $s^2 = 1$. A component -based score, on the other hand, is a linear composite of the variables that demonstrated meaningful loadings. It is obtained by simply summing the response of the variables in each component with significant loadings or by taking a variable with the highest weight.

3.3.2. Nonlinear principal component analysis

3.3.2.1. Background

Nonlinear principal component analysis, also known as categorical principal components analysis (CATPCA), is appropriate for data reduction when variables are categorical (e.g. nominal variables - consist of unordered categories, or ordinal variables - consist of unordered categories), and the researcher is concerned with identifying the underlying components of a set of variables while maximizing the amount of variance accounted for in those items (by the principal components).

In nonlinear PCA, categories of such variables are assigned numeric values through a process called optimal quantification (also referred to as optimal scaling, or optimal scoring). Such numeric values are referred to as category quantifications; the category quantifications for one variable together form that variable's transformation. Optimal quantification replaces the category labels with category quantifications in such a way that as much as possible of the variance in the quantified variables is accounted for. Nonlinear PCA achieves the very same objective as linear PCA for quantified categorical variables. If all variables in nonlinear PCA are numeric, the nonlinear PCA and linear PCA solution are exactly equal, because in that case no optimal quantification is required, and the variables are merely standardized.

Correlations are not computed between the observed variables, but between the quantified variables. As opposed to the correlation matrix in linear PCA, the correlation matrix in nonlinear PCA is not fixed; rather, it is dependent on the type of quantification, called an analysis level that is chosen for each of the variables. In contrast to the linear PCA solution,

the nonlinear PCA solution is not derived from the correlation matrix, but iteratively computed from the data itself, using the optimal scaling process to quantify the variables according to their analysis level. The objective of optimal scaling is to optimize the properties of the correlation matrix of the quantified variables. Specifically, the method maximizes the first p eigenvalues of the correlation matrix of the quantified variables, where p indicates the number of components that are chosen in the analysis.

3.3.2.2. A nonlinear PCA and linear PCA: Similarities and Differences

Both methods provide eigenvalues, component loadings, and component scores. In both cases, the eigenvalues are overall summary measures that indicate the variance accounted for by each component. Component loadings are measures obtained for the variables, and in both linear and nonlinear PCA, are equal to a Pearson correlation between the principal component and either an observed variable (linear PCA) or a quantified variable (nonlinear PCA). If nonlinear relationships between variables exist, and nominal or ordinal analysis levels are specified, nonlinear PCA leads to a higher variance accounted for than linear PCA, because it allows for nonlinear transformations [20]**.**

The principal components in linear PCA are weighted sums of the original variables, whereas in nonlinear PCA they are weighted sums of the quantified variables. In both methods the components consist of standardized scores.

The difference is that in linear PCA the measured variables are directly analysed, while in nonlinear PCA the measured variables are quantified during the analysis. Another difference is the nestedness of the solution.

3.3.2.3. Nestedness of the components solution

Linear PCA maximizes the variance accounted for (VAF) of the first component over linear transformations of the variables, and then maximizes the VAF of the second component that is orthogonal to the first, and so on. This is sometimes called consecutive maximization. Linear PCA also maximizes the total VAF in p dimensions simultaneously by projecting the original variables from a q-dimensional space onto a p-dimensional component space. This is called simultaneous maximization. Consecutive maximization of the VAF in p components is identical to simultaneous maximization, and we say that linear PCA solutions are nested for different values of p [20].

In nonlinear PCA, consecutive and simultaneous maximization will give different results. Nonlinear PCA maximizes the VAF of the first p components simultaneously over nonlinear transformations of the variables. The eigenvalues are obtained from the correlation matrix among the quantified variables, and the sum of the first p eigenvalues is maximized. In this case, the solutions are usually not nested for different values of p [20].

We employed principal component analysis to 27 variables measuring quality of life of breast cancer patients. Our main aim is to reduce the 27 variables into a single variable which will account for as much of the variance in the data set as possible. The component obtained was named quality of life. We used the median score as the cut off criterion, patients whose score were above the median were considered to have high quality of life while those with median and below score were considered to have low quality of life. The quality of life will therefore have two outcomes and it was used as a response variable in the binary logistic regression.
The independent variables are socio-demographic and clinical characteristics of the patients. The socio-demographic characteristics are age (years), education, marital status and parity. Clinical characteristics of the patients are type of treatment received and stage of the disease.

3.3.3. Logistic regression

3.3.3.1. Introduction

An ordinary least squares regression model is of the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad (3.3.0)$$

Ordinary least squares regression model however is restrictive in the sense that:

- They only cater for continuous response variables having normal distribution
- The relationship between the response and the predictors is a simple ("identity") function

The generalized linear models go beyond this in two major aspects:

- The response variables can have a distribution other than normal; i.e. any distribution within a class of distributions known as "exponential family of distributions".
- Instead of having $\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$ we can allow for transformations:

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

The distribution of the response variable is extended to the exponential family of distributions. For any random variable Y with probability distribution function $f(y,\theta)$; where θ is an unknown parameter; the probability distribution function can be expressed in the form:

$$f(y,\theta) = exp[a(y)b(\theta) + c(\theta) + d(y)],$$

then the distribution is said to be belong to the exponential family of distributions.

Further if $(y) = y$, then the distribution is said to be in canonical form and $b(\theta)$ is known as natural parameter.

Example; Binomial distribution

$Y \sim b(Sampe = n, probability\ of\ success = \theta)$

$$f(y,\theta) = \binom{n}{y}\theta^{y}(1-\theta)^{n-y},$$

where n is the sample size, y is the random variable, θ is the probability of success, $(1-\theta)$ is the probability of failure.

$f(y,\theta)$ can be written as:

$$f(y,\theta) = exp\left[\ln\binom{n}{y} + yln\theta + (n-y)ln(1-\theta)\right]$$

Thus

$$= exp\left[\ln\binom{n}{y} + yln\theta + nln(1-\theta) - yln(1-\theta)\right]$$

$$= exp\left[\ln\binom{n}{y} + yln\left(\frac{\theta}{1-\theta}\right) + nln(1-\theta)\right]$$

On rearranging, we get

$$= exp\left[yln\left(\frac{\theta}{1-\theta}\right) + nln(1-\theta) + \ln\binom{n}{y}\right]$$

where $a(y) = y; b(\theta) = ln\left(\frac{\theta}{1-\theta}\right); c(\theta) = nln(1-\theta); d(y) = \ln\binom{n}{y}$

This shows that a binomial distribution belongs to exponential family of distributions that are in canonical form, with the natural parameter of $ln\left(\frac{\theta}{1-\theta}\right)$ and canonical link function called logit.

For exponential family of distributions in canonical form, the natural parameter is used as the link function. It is known as canonical link function.

Logistic regression is a model in which the outcome is measured on a binary scale. For example, the responses may be success and failure. We can define a binary variable as;

$$Y = \begin{cases} 1, & Low\ quality\ of\ life \\ 0, & Better\ quality\ of\ life \end{cases}$$

Assumptions of logistic regression

- Logistic regression does not assume linear relationship between the dependent and the independent variables.
- Dependent variable must have two categories for binary logistic regression and more than two for multinomial logistic regression.
- The categories must be mutually exclusive and exhaustive; an individual must belong to only one group and every case must be a member of one of the groups.
- Large samples are needed than for linear regression because maximum likelihood coefficients are large sample estimates. A minimum of 50 cases per predictor is recommended.

3.3.3.2. Logistic regression model

Suppose we have only one predictor *x*, then for a sample size of n, logistic regression model is called simple logistic regression model and is given as;

$$logit(p_i) = ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i \qquad (3.31)$$

where x_i is the predictor variable, β_i is the unknown regression coefficient which will be estimated, p_i is the probability of success and $1 - p_i$ is the probability of failure.

The model is equivalent to:

$$\frac{p_i}{1-p_i} = \exp(\beta_0 + \beta_1 x_i)$$

$$p_i = (1 - p_i)\exp(\beta_0 + \beta_1 x_i)$$

$$p_i\,(1 + \exp(\beta_0 + \beta_1 x_i) = \exp(\beta_0 + \beta_1 x_i)$$

And finally, p_i can be written as;

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1+\exp(\beta_0 + \beta_1 x_i)} \qquad (3.32)$$

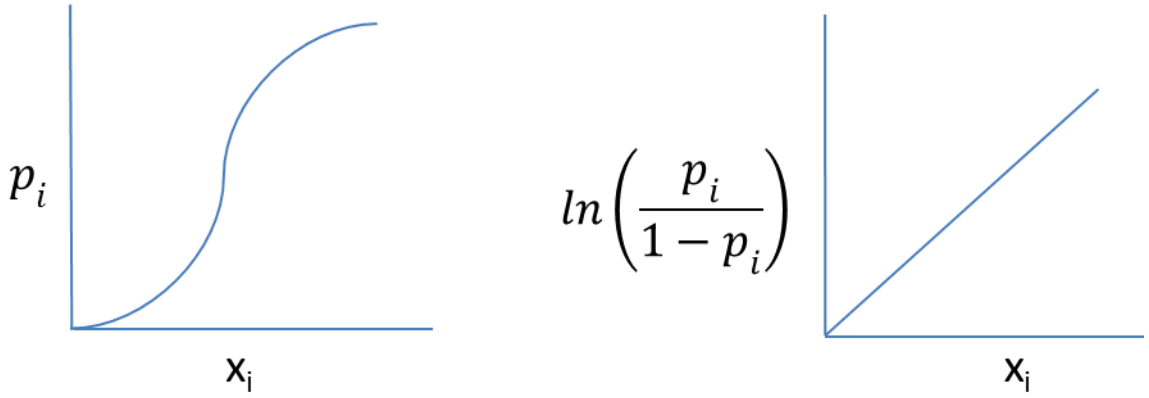

Figure 5. Logistic regression models: logistic curve (left) and linear (right).

To get a linear model, we use the log of the odds of an event (equation 3.31) occurring, where;

$$odds\ of\ event = \frac{p}{1-p}.$$

The predictor variable x_i can either be continuous or categorical. Categorical variables are categorized in dummy variables.

If more than one predictor variable is used to explain the response variable then the model becomes;

$$logit(p_i) = ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}\,, \quad i = 1,2,\ldots\ldots,n \qquad (3.33)$$

or $\quad ln\left(\frac{p_i}{1-p_i}\right) = \sum_{k=0}^{k} \beta_k x_{ik}, \qquad (3.34)$

where $(x_1, x_2, \ldots, x_n)$ are the predictor variables, and $(\beta_0, \beta_1, \ldots, \beta_k)$ are the unknown regression coefficients.

3.3.3.3. Interpretation of parameters

If the predictor **(X)** is a continuous variable; for any two values of predictor X = t and X = t + 1 where t is a constant;

$$(odds\ when\ X = t) = \frac{p_i}{1-p_i} = e^{(\beta_0 + \beta_1 t)},$$

and

$$(odds\ when\ X = t+1) = \frac{p_i}{1-p_i} = e^{(\beta_0 + \beta_1 (t+1))}.$$

Thus:

$$odds\ ratio\ (OR) = \frac{odds\ when\ X=t+1}{odds\ when\ X=t} = \frac{e^{\beta_0} e^{\beta_1 t} e^{\beta_1}}{e^{\beta_0} e^{\beta_1 t}} = e^{\beta_1}$$

e^{β_1} is the change in likelihood of event occurring for every additional measure of the predictor variable.

If $\beta_1 = 0$, then $e^{\beta_1} = 1$. The implication of this is that the predictor variable is not significant in predicting the response variable. Another interpretation is that if odds ratio is 1, then the odds of the event occurring for both groups is the same. Thus the value 1 for odds ratio is used as a reference point for interpretation of odds ratio.

If $\beta_1 < 0$, then $e^{\beta_1} < 1$. The event is e^{β_1} times less likely to occur for every unit increase in predictor. Alternatively, the event is $100(1 - e^{\beta_1})\%$ less likely to occur for every unit increase in predictor.

If $\beta_1 > 0$, then $e^{\beta_1} > 1$. The event is e^{β_1} times more likely to occur for every unit increase in predictor. Alternatively, if $1 < e^{\beta_1} < 2$ the event is $100(e^{\beta_1} - 1)\%$ more likely to occur for every unit increase in predictor.

If the predictor is a categorical variable, first select one level of the variable as a reference group and then create a dummy variable. For example;

X = 0 and for the other level X = 1

$(odds\ when\ X = 0) = \frac{p_i}{1-p_i} = e^{(\beta_0 + \beta_1 0)} = e^{\beta_0}$

and

$(odds\ when\ X = 1) = \frac{p_i}{1-p_i} = e^{(\beta_0 + \beta_1 1)}$.

Thus;

$odds\ ratio\ (O.R) = \frac{odds\ when\ X=1}{odds\ when\ X=0} = \frac{e^{\beta_0} e^{\beta_1 1}}{e^{\beta_0}} = e^{\beta_1}$

If $\beta_1 = 0$, then $e^{\beta_1} = 1$ The implication of this is that the predictor variable is not significant in predicting the response variable. Another interpretation is that if odds ratio is 1, then the odds of the event occurring for both groups is the same. Thus the value 1 for odds ratio used as a reference point for interpretation of odds ratio.

If $\beta_1 < 0$, the event is e^{β_1} times less likely to occur for other group compared to the reference group. Alternatively, the event is $100(1 - e^{\beta_1})\%$ less likely to occur for other group compared to the reference group.

If $\beta_1 > 0$, the event is e^{β_1} times more likely to occur for other group compared to the reference group. Alternatively, if $1 < e^{\beta_1} < 2$ the event is $100(e^{\beta_1} - 1)\%$ more likely to occur for every unit increase in predictor.

3.3.3.4. Estimation of regression parameters

The goal of logistic regression is to estimate k +1 unknown parameters in equation 3.34. This is done with maximum likelihood estimation which entails finding the set of parameters for which the probability of the observed data is greatest. The maximum likelihood equation is derived from the probability distribution of the dependent variable. Since each y_i represents a binomial count in the ith population, the joint probability density function of y is:

$$f(y|\beta) = \prod_{i=1}^{N} \frac{n_i!}{y_i!(n_i - y_i)!} p_i^{\ y_i} (1 - p_i)^{n_i - y_i} \qquad (3.35)$$

The likelihood function expresses the values of β in terms of known, fixed values for y. Thus,

$$L(\beta|y) = \prod_{i=1}^{N} \frac{n_i!}{y_i!(n_i-y_i)!} p_i^{\,y_i}(1-p_i)^{n_i-y_i} \quad (3.36)$$

The maximum likelihood estimates are the values for β that maximize the likelihood function in equation 3.36. The critical points of a function (maxima and minima) occur when the first derivative equals 0. If the second derivative evaluated at that point is less than zero, then the critical point is a maximum.

Attempting to take the derivative of equation 3.36 with respect to β is a difficult task due to the complexity of multiplicative terms. Fortunately, the likelihood equation can be considerably simplified. This is because the factorial terms do not contain any of the p_i and as a result, they are essentially constants that can be ignored. Also, since$(a)^{x-y} = \frac{a^x}{a^y}$, and after rearranging terms, the equation to be maximized can be written as:

$$\prod_{i=1}^{N} \left(\frac{p_i}{(1-p_i)}\right)^{y_i} (1-p_i)^{n_i} \quad (3.37)$$

Equation 3.32 for multivariate data can be written as:$p_i = \left(\frac{e^{\sum_{k=0}^{k}\beta_k x_{ik}}}{1+e^{\sum_{k=0}^{k}\beta_k x_{ik}}}\right)$

(3.38)

Substituting equation 3.38 for equation 3.37 becomes:

$$\prod_{i=1}^{N} \left(e^{\sum_{k=0}^{k}\beta_k x_{ik}}\right)^{y_i} \left(1 - \frac{e^{\sum_{k=0}^{k}\beta_k x_{ik}}}{1+e^{\sum_{k=0}^{k}\beta_k x_{ik}}}\right)^{n_i} \quad (3.39)$$

Using $(a^x)^y = a^{xy}$ to simplify the first product and replacing 1 with $\frac{1+e^{\sum_{k=0}^{k}\beta_k x_{ik}}}{1+e^{\sum_{k=0}^{k}\beta_k x_{ik}}}$ to simplify the second product. Equation 3.39 can now be written as:

$$\prod_{i=1}^{N} e^{y_i \sum_{k=0}^{k}\beta_k x_{ik}} \left(1 + e^{\sum_{k=0}^{k}\beta_k x_{ik}}\right)^{-n_i} \quad (3.40)$$

This is the kernel of the likelihood function to maximize. However, it is still cumbersome to differentiate and can be simplified a great deal further by taking its log. Since the logarithm is a monotonic function, any maximum of the likelihood function will also be a maximum of the log likelihood function and vice versa. Thus, taking the natural log of equation 3.40 yields the log likelihood function:

$$l(\beta) = \sum_{i=1}^{N} y_i\left(\sum_{k=0}^{k}\beta_k x_{ik}\right) - n_i ln\left(1 + e^{\sum_{k=0}^{k}\beta_k x_{ik}}\right) \quad (3.41)$$

To find the critical points of the log likelihood function, set the first derivative with respect to each β equal to zero.

In differentiating equation 3.41 with respect to each β,

$$\frac{\partial l(\beta)}{\partial \beta_k} = \sum_{i=1}^{N} y_i x_{ik} - n_i \cdot \frac{1}{1+e^{\sum_{k=0}^{k} \beta_k x_{ik}}} \cdot \frac{\partial\left(1+e^{\sum_{k=0}^{k} \beta_k x_{ik}}\right)}{\partial \beta_k}$$

$$\frac{\partial l(\beta)}{\partial \beta_k} = \sum_{i=1}^{N} y_i x_{ik} - n_i \cdot \frac{1}{1+e^{\sum_{k=0}^{k} \beta_k x_{ik}}} \cdot e^{\sum_{k=0}^{k} \beta_k x_{ik}} \cdot x_{ik} \quad (3.42)$$

Substituting equation 3.38 for equation 3.42 becomes;

$$\frac{\partial l(\beta)}{\partial \beta_k} = \sum_{i=1}^{N} y_i x_{ik} - n_i \cdot p_i \cdot x_{ik} \quad (3.43)$$

The maximum likelihood estimates for β can be found by setting each of the k+1 equations in equation 3.43 equal to zero and solving for each β_k.

The critical point will be a maximum if the second partial derivatives is negative definite.

$$\frac{\partial^2 l(\beta)}{\partial \beta_k \partial \beta_{k'}} = \frac{\partial}{\partial \beta_{k'}} \sum_{i=1}^{N} y_i x_{ik} - n_i \cdot p_i \cdot x_{ik}$$

$$\frac{\partial^2 l(\beta)}{\partial \beta_k \partial \beta_{k'}} = -\sum_{i=1}^{N} n_i x_{ik} \frac{\partial}{\partial \beta_{k'}} \left(\frac{e^{\sum_{k=0}^{k} \beta_k x_{ik}}}{1+e^{\sum_{k=0}^{k} \beta_k x_{ik}}} \right)$$

$$\frac{\partial^2 l(\beta)}{\partial \beta_k \partial \beta_{k'}} = -\sum_{i=1}^{N} n_i x_{ik} p_i (1-p_i) x_{ik'} \quad (3.44)$$

3.3.3.5. Test of goodness of model fit

One way of assessing the adequacy of model fitted is to compare it with another model. The model commonly used for comparison is a more general model with the maximum number of parameters that can be estimated called saturated model. Using the saturated model; the hypotheses are:

H_0: the fitted model is better fit vs. H_1: the saturated model is a better fit

The test statistic for this test is the log-likelihood ratio test statistic also known as deviance statistic (D)**.** It is calculated by subtracting the log-likelihood of fitted model by the saturated model.

$$D = 2\left[In\ L_s - In\ L_f\right] = -2\left[InL_f\ - InL_s\right]$$

$In\ L_f$ -Maximum of the log likelihood of the fitted model

$In\ L_s$ - Maximum achievable log likelihood of a saturated model

The statistics has an approximate chi square distribution with n-p degrees of freedom;

$$D \sim \chi^2{}_{n-p}$$

where **n** is the number of observations and **p** is the number of parameters.

Small values of D indicate a good fit, i.e. we fail to reject the null hypothesis.

3.3.3.6. Statistical inference on regression parameters and odds ratio estimation

The $100(1-\propto)\%$ confidence interval for β_j and OR are;

β_j is $\hat{\beta}_j \pm Z_{\frac{\propto}{2}} s.e(\hat{\beta}_j)$

$OR = e^{\beta_j}$ is $\exp\left(\hat{\beta}_{Jj} \pm Z_{\frac{\propto}{2}} s.e\left(\hat{\beta}_j\right)\right)$

Hypothesis testing:

For $\boldsymbol{\beta_j}$

H_0: $\beta_j = 0$ *versus* H_1: $\beta_j \neq 0$

Test statistics: Z score which is given as; $Z = \frac{\hat{\beta}_j}{s.e\,(\hat{\beta}_j)} \sim N(0,1)$

Wald's statistic $Z^2 = \frac{\hat{\beta}_j{}^2}{Var\,(\hat{\beta}_1)} \sim \chi^2(1)$

Rule: reject H_0 if the p-value is less than $\propto$ level of significance.

For OR

H_0: $OR = 1$ verse *versus* H_1: $OR \neq 1$

Rule: Reject H_0 at $\propto$ level of significance if the 100(1- $\propto$) % confidence interval for OR does not contain 1.

4. DATA ANALYSIS AND RESULTS

This chapter discusses data analysis, principal component analysis and logistic regression results.

4.1.Data analysis

Responses to 27 items of EORTC QLQ-C30 questionnaire were subjected to principal component analysis. One variable which had a variance of zero (need help with eating, dressing, washing yourself or using toilet) was not included in the analysis. We specified an ordinal level analysis for all the 27 variables. Symptoms scale and items were reversed prior to analysis. We restricted the PCA analysis to one component solution and the obtained component was used as a response variable in logistic regression.

A significance level of 5% and two-tail test were used. SPSS version 2.1 was used to analyze the data.

4.2. Nonlinear principal component analysis

4.2.1. Assessing the suitability of linear principal component analysis

Before any analysis was done, assumption of linearity was assessed using a scatterplot with a trend line to determine whether linear or nonlinear principal component analysis will be used. The scatterplot of the few variables that were picked did not follow a straight line implying that the relationship between any two variables is not linear (Figure 6). Presence of collinearity was also assessed using the correlation matrix (see Appendix 1). Majority of the variables had a correlation coefficient of > 0.3. Because of lack of linearity, we employed nonlinear principal component analysis.

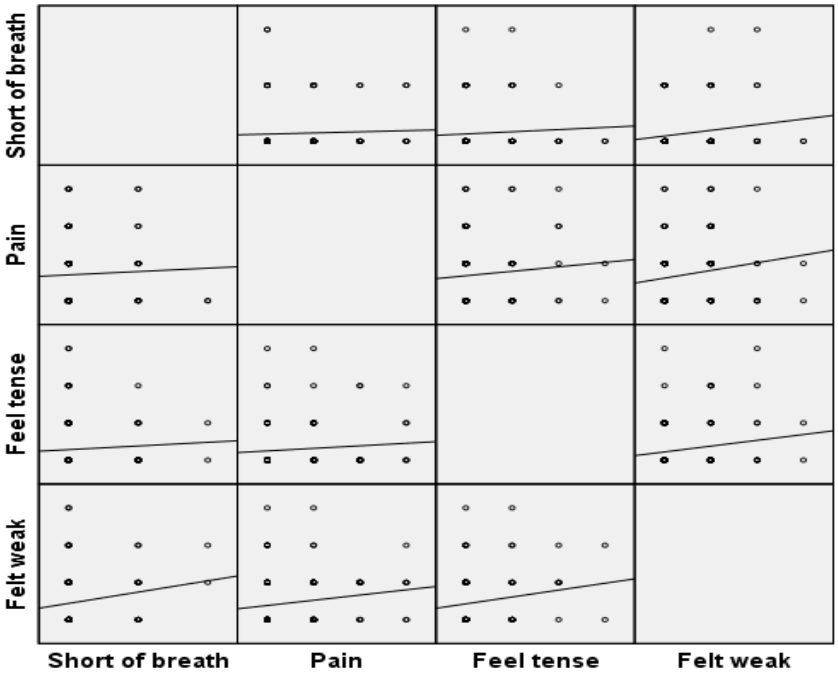

Figure 6: Matrix scatterplot

4.2.2. Choosing the number of components

We performed CATPCA analysis on quality of life items with one component. The total eigenvalue were 6.622 and PVAF obtained was 23.7%. We further inspected the component loadings (Table 1), and variables with low loadings (< 0.5) were dropped because they were not contributing substantially to the principal component. The results obtained when CATPCA analysis was performed on the remaining variables resulted in higher total PVAF of 46.8% compared to the previous 23.7%. Table 2 represents component loadings on one-component solution after items with low loadings were dropped.

Table 1. Component loadings of one-component solution

Items	**Loadings**	**Items**	**Loadings**
Limited in doing daily activities	0.795	Having vomited	-0.125
Having trouble taking long walk	0.716	Having nauseated	-0.226
Need to stay in bed/chair during the day	0.683	Lack of appetite	-0.295
Have trouble doing strenuous activity	0.641	Constipation	-0.295
Limited in pursuing your hobbies	0.640	Had Diarrhoea	-0.327

Having trouble taking short walk	0.612	Short of breath	-0.334
Physical condition or medical condition interferes with social activities	0.491	Fatigue	-0.444
Difficulty in remembering things	0.449	Trouble sleeping	-0.514
Feel depressed	0.445	Felt weak	-0.543
Feel tense	0.436	Pain	-0.568
Difficulty in concentration on things like reading a newspaper	0.425	Pain interfere with daily activities	-0.660
Feel worry	0.359	Need of rest	-0.671
Physical condition or medical condition interferes with family life	0.347		
Feel irritable	0.276		
Financial difficulties	0.251		

Table 2. Component loadings of one-component solution after variables with low loadings dropped

Items	Loadings
Limited in doing daily activities	0.684
Having trouble taking long walk	0.790
Need to stay in bed/chair during the day	0.678
Limited in pursuing your hobbies	0.700
Have trouble doing strenuous activity	0.867
Having trouble taking short walk	0.679
Trouble sleeping	-0.689
Felt weak	-0.612
Pain	-0.699
Pain interfere with daily activities	-0.527
Need of rest	-0.527

Eleven items were used in calculation of component based scores because of their high correlation with the first component. The items comprise of all the physical and role functioning scale, all items of pain scale, insomnia and 2 items of the fatigue scale. Component based scores were obtained by calculating the average of all the 11 items. The new variable was named ‘quality of life’, and was used as a response variable in logistic regression.

4.2.3. Hypothesized model

Having obtained a single variable called ‘quality of life’ from nonlinear principal component analysis; the model used in logistic regression is thus given by;

$$ln\left(\frac{p_i\ (Low\ quality\ of\ life)}{1-p_i\ (Low\ quality\ of\ life)}\right) = \beta_0 + \beta_1 Age + \beta_2 married + \beta_3\ primary\ education\ level + \beta_4\ secondary\ education\ level + \beta_7\ more\ than\ 4\ children + \beta_{11}\ late\ stage\ of\ disease + \beta_8\ \text{Tamoxifen} + \beta_9\ \text{Radiotherapy}$$

4.3. Logistic regression analysis

This section presents the descriptive analysis and logistic regression results.

4.3.1. Descriptive analysis

The mean age of the respondents was 49.4 ±10.2 years. More than two third of the respondents were married and majority (90.9%) had secondary education level and below. More than half of the respondents were in their late stages of the disease. At the time of interview patients were exposed to different stages of treatment therapy sequence. Thirty eight percent of the respondents were on chemotherapy, 27.5% on radiotherapy, 20.4% on tamoxifen and 14.1% were on surgery. Majority of the respondents had undergone surgery before the interview (Table 3).

Table 3. Descriptive analysis

Variable group	Variables	Overall	Better QOL	Low QOL
Age in years	Mean(SD)	49.4(10.2)	50.3 (10.2)	49.0 (10.3)
Marital status	Married	96 (67.6)	33 (34.4)	63 (65.6)
	Others	46 (32.4)	16 (34.8)	30 (65.2)
Education	Primary and none	62 (43.7)	23 (37.1)	39 (62.9)
	Secondary level	67 (47.2)	22 (32.8)	45 (67.2)
	Tertiary level	13 (9.2)	4 (30.8)	9 (69.2)
Parity	0 – 4	108 (76.6)	35 (32.4)	73 (67.6)
	More than 4	33 (23.4)	14 (42.4)	19 (57.6)
Stage	Early Stage	47 (38.2)	16 (34.0)	31 (66.0)
	Late Stage	76 (61.8)	27 (35.5)	49 (64.5)
Treatment	Surgery	20 (14.1)	2 (10.0)	18 (90.0)
	Tamoxifen	29 (20.4)	3 (10.3)	26 (89.7)
	Radiotherapy	39 (27.5)	17 (43.6)	22 (56.4)
	Chemotherapy	54 (38.0)	27 (50.0)	27 (50.0)
Total			49 (34.5)	93 (65.5)

The median quality of life score of the respondents was 2.45 with a minimum of 1.73 and maximum of 3.09. Using median as the cut off, 34.5% were above the median score and were considered to have better quality of life, while 65.5% were considered to have poor quality of life.

4.3.2. Logistic regression

4.3.2.1. Model fit and likelihood function

The likelihood ratio (-2LL) is a test of significance between the likelihood ratio for the fitted model minus the likelihood ratio for the null model. The smaller the -2LL value the better the fit. The value obtained was 136.643 which is fairly small and thus the model is a good fit (Table 5).

Omnibus Tests of Model Coefficients reports the chi square associated with each step in a stepwise model (Table 4). All the values are the same since there is only one step from the null model to the block containing predictors. The test is based on the null hypothesis that the null model is a better fit verse the alternative hypothesis that states fitted model is a better fit. The p value obtained was 0.013 which is less than 0.05. It therefore implies that the fitted model is significantly different from the null model, thus a better fit.

Table 4. Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	20.828	9	0.013
	Block	20.828	9	0.013
	Model	20.828	9	0.013

Model Summary: The model summary (Table 5) provides some **R** estimation. Nagelkerke R Square is preferred because it ranges from 0 to 1. Nagelkerke R Square was 0.218, which is a bit minimal. This indicates 21.8% relationship between the predictors and the prediction.

Table 5. Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	136.643	0.158	0.218

4.3.2.2. Logistic regression estimate, Wald test and odds ratio

Table 6. Logistic regression estimate, Wald test, odds ratio and 95% CI for odds ratio

Variable	**Variable category**	$\boldsymbol{\beta}$	**S.E.**	**Wald test**	**df**	**Sig.**	**Exp** (β)	**95% CI for Exp ($\boldsymbol{\beta}$)**	
								Lower	**Upper**
Age	Age (years)	-0.019	0.032	0.346	1	0.556	0.981	0.922	1.045
Marital status	Married	-0.023	0.502	0.002	1	0.963	0.977	0.365	2.614
Education	Tertiary			1.843	2	0.398			
	Primary and none	1.324	0.996	1.767	1	0.184	3.759	0.533	26.490
	Secondary	1.105	0.880	1.577	1	0.209	3.019	0.538	16.929
Parity	>4 children	-0.502	0.609	0.678	1	0.410	0.606	0.183	1.999
Stage	Late stage	0.195	0.470	0.172	1	0.678	1.215	0.484	3.052
Treatment	Chemotherapy			13.205	3	0.004			
	Surgery	2.172	0.846	6.595	1	0.010	8.778	1.673	46.070
	Tamoxifen	2.097	0.704	8.859	1	0.003	8.141	2.046	32.385
	Radiotherapy	0.253	0.505	0.250	1	0.617	1.288	0.478	3.466
	Constant	-0.203	1.617	0.016	1	0.900	0.816		

From the Table 6, the estimated model is;

$$ln\left(\frac{p_i\,(Low\ quality\ of\ life)}{1-p_i\,(Low\ quality\ of\ life)}\right)$$
$$= -0.203 - 0.019\,Age - 0.023\,married$$
$$+ 1.105\,secondary\ education\ level + 1.324\,primary\ and\ none$$
$$- 0.502\ \ more\ than\ 4\ children\ + 0.195\,late\ stage + 2.172\,surgery$$
$$+ 2.097\ \text{tamoxifen}\ +\ 0.253\ \ radiotherapy$$

Significance of the predictors

The Wald test shows that surgery and tamoxifen treatments are statistically significant predictors since the p value; 0.010 and 0.001 respectively are all less than 0.05. This is also supported by their 95% CI for OR because they do not contain 1. Age, marital status, education, parity, stage of the disease and treatment - radiotherapy are not significant predictors.

OR interpretation

Age: OR was 0.981. This means that for every additional age the respondents are 1.9% less likely to have a poor quality of life.

Marital status: OR was 0.977. This means that married respondents are 2.3% less likely to have poor quality of life compared to the single, widowed or divorced patient.

Education: OR for respondents with secondary education was 3.019 and for respondents with below secondary education was 3.759. This implies that respondents with secondary education are 3.019 times more likely to have poor quality of life while those on primary or no education are 3.759 times more likely to have poor quality of life compared to those with tertiary education.

Parity: Respondents with more than 4 children were 39.4 % less likely to have poor quality of life compared to those with four and below children.

Stage: Respondents in late stage of the disease were 21.5% more likely to have poor quality of life compared to those in the early stages.

Treatment: Respondents on surgery, tamoxifen and radiotherapy were more likely to have poor quality of life compared to those on chemotherapy. Specifically, respondents receiving surgery and tamoxifen were more than 8 times while those receiving radiotherapy were 1.3 times more likely to have poor quality of life.

5. CONCLUSIONS AND RECOMMENDATIONS

This chapter presents the conclusion and recommendation

In this study, we set out to identify a single variable that indicates the health related quality of life of breast cancer patients and to find out how this is influenced by patients' socio-demographics characteristics, stage of the disease and treatment received. Our sample population was patients receiving cancer treatments at the Haemato-oncology and Cancer Treatment Centre of Kenyatta National Hospital. One hundred and forty two patients were recruited in the study. The study population was fairly young and more than half of the respondents had secondary education and above. More than half of the respondents were in their late stage of the disease and were exposed to different stages of treatment therapy.

The variable obtained (QOL) comprises of all the physical and role functioning scale, pain scale, insomnia and 2 items of the fatigue scale. It implies that physical, role, pain, fatigue and insomnia are associated with each other. This result is in agreement with [12] which found that fatigue is associated with pain and insomnia. The results are also in line with Kuwait study [13] which found that physical and role functioning are highly correlated.

The median quality of life of the respondents was 2.45, the lowest score being 1.73 while the highest was 3.09. Using median as the cut off, more than half of the respondents were classified as having poor quality of life. The possible reasons for this is that most of the respondents were in their late stage of the disease and limited resources in the healthcare centers.

Results show that surgery and tamoxifen treatments are statistically significant predictors of quality of life of the respondents while age, marital status, education, parity, stage of the disease and treatment -radiotherapy are not significant predictors.

Surgery and tamoxifen treatments are significant predictors while radiotherapy is not. These results are in line with the study done in USA [17] which revealed treatment received was not significant predictor of fatigue. The findings from the odds of having poor quality of life showed that respondents receiving surgery, tamoxifen and radiotherapy treatments were more likely to have poor quality of life compared to those on chemotherapy.

Even though our findings revealed that respondents in late stage of the disease are more likely to have poor quality of life, stage of the disease was not a significant predictor. This

finding disagrees with a study done in India [14] which found that stage of the disease is significant indicator of respondents' quality of life. The possible reason for the two different results is that in our study, stage of the disease of 19 respondents was not documented and they were not included in the analysis.

All the demographic characteristics (age, education level, marital status and parity) used in the study were not significant predictors of respondents' quality of life. A study done in Brazil [12] found that age is predictive factor of fatigue which contradicts with our findings. However, it agrees with our other findings that education level, marital status and parity are not related with fatigue. Even though the selected respondents' demographic characteristics are not significant predictors, we sought to find out their odds of having poor quality of life. For every additional age, respondents are 1.9% less likely to have poor quality of life. Married respondents are 2.3% less likely to have poor quality of life compared to those who are not married. Respondents who have less education are 3 times more likely to have poor quality of life compared to those with more education. Finally, respondents with more than four children are 39.4% less likely to have poor quality of life. These findings diverge somewhat from other studies that have shown age is significant predictors of respondents' quality of life. The possible reasons may be difference in measurements techniques and difference in socio-demographic characteristics of the samples assessed, as well as geographic locations from which the samples were drawn.

Overall, our findings indicate that breast cancer patients have poor quality of life and socio-demographics and stage of the disease are not significant predictors of patients' quality of life. Even though stage of the disease is not significant predictor, it was noted that patients in their late stage of the disease are more likely to have poor quality of life compared to those in the early stage. Treatment seems to play a role in determining the patients' quality of life because patients receiving surgery and taxomifen reported lower quality of life score. However, it was reported that radiotherapy was not a significant predictor of patients' quality of life.

The limitation of our study is the study design. The study was conducted at Kenyatta National Hospital, and the findings cannot be generalized to the entire breast patient in Kenya. Despite the limitation, this study may be used in management of quality of life in breast cancer patients by directing innovative interventions that improve quality of life of the patients. For

example, improvement of facilities in the healthcare centres so that patients receive quality treatment and promotion of activities that improve early screening of breast cancer.

References

[1] "MedicineNet - Health and Medical Information Produced by Doctors." [Online]. Available: http://www.medicinenet.com/script/main/hp.asp. [Accessed: 10-Jun-2014].

[2] "Latest world cancer statistics - Global cancer burden rises to 14.1 million new cases in 2012: Marked increase in breast cancers must be addressed . Estimated Incidence, Mortality and Prevalence Worldwide in 2012 by International Agency for Research on Cancer (World Health Organization) | Ronning Against Cancer to run against cancer, to support cancer charities and people." .

[3] F. Bray, J.-S. Ren, E. Masuyer, and J. Ferlay, "Global estimates of cancer prevalence for 27 sites in the adult population in 2008," *Int. J. Cancer*, vol. 132, no. 5, pp. 1133–1145, Mar. 2013.

[4] L. Van Esch, B. L. Den Oudsten, and J. De Vries, "The World health Organization quality of life instrument-Short form (WHOQOL-BREF) in women with breast problems," *Int. J. Clin. Health Psychol.*, vol. 11, no. 1, pp. 5–22, 2011.

[5] N. K. Aaronson, S. Ahmedzai, B. Bergman, M. Bullinger, A. Cull, N. J. Duez, A. Filiberti, H. Flechtner, S. B. Fleishman, J. C. J. M. d. Haes, S. Kaasa, M. Klee, D. Osoba, D. Razavi, P. B. Rofe, S. Schraub, K. Sneeuw, M. Sullivan, and F. Takeda, "The European Organization for Research and Treatment of Cancer QLQ-C30: A Quality-of-Life Instrument for Use in International Clinical Trials in Oncology," *JNCI J. Natl. Cancer Inst.*, vol. 85, no. 5, pp. 365–376, Mar. 1993.

[6] P. O. S. ABUSE, "PROGRAMME ON MENTAL HEALTH," 1997.

[7] A. G. Pallis and I. A. Mouzas, "Instruments for quality of life assessment in patients with gastrointestinal cancer," *Anticancer Res.*, vol. 24, no. 3B, pp. 2117–2121, 2004.

[8] W. O. Spitzer, A. J. Dobson, J. Hall, E. Chesterman, J. Levi, R. Shepherd, R. N. Battista, and B. R. Catchlove, "Measuring the quality of life of cancer patients: a concise QL-index for use by physicians," *J. Chronic Dis.*, vol. 34, no. 12, pp. 585–597, 1981.

[9] J. C. de Haes, F. C. van Knippenberg, and J. P. Neijt, "Measuring psychological and physical distress in cancer patients: structure and application of the Rotterdam Symptom Checklist.," *Br. J. Cancer*, vol. 62, no. 6, pp. 1034–1038, Dec. 1990.

[10] D. F. Cella, D. S. Tulsky, G. Gray, B. Sarafian, E. Linn, A. Bonomi, M. Silberman, S. B. Yellen, P. Winicour, and J. Brannon, "The Functional Assessment of Cancer Therapy scale: development and validation of the general measure.," *J. Clin. Oncol.*, vol. 11, no. 3, pp. 570–579, Mar. 1993.

[11] A. Thompson, K. Brennan, A. Cox, J. Gee, D. Harcourt, A. Harris, M. Harvie, I. Holen, A. Howell, R. Nicholson, M. Steel, C. Streuli, and $author firstName $author.lastName, "Evaluation of the current knowledge limitations in breast cancer research: a gap analysis," *Breast Cancer Res.*, vol. 10, no. 2, p. R26, Mar. 2008.

[12] A. C. G. Cavalli Kluthcovsky, A. A. Urbanetz, D. S. Carvalho, E. M. C. Pereira Maluf, G. C. Schlickmann Sylvestre, and S. B. Bonatto Hatschbach, "Fatigue after treatment in breast cancer survivors: prevalence, determinants and impact on health-related quality of life," *Support. Care Cancer*, vol. 20, no. 8, pp. 1901–1909, Aug. 2012.

[13] S. A. Alawadi and J. U. Ohaeri, "Health – related quality of life of Kuwaiti women with breast cancer: a comparative study using the EORTC Quality of Life Questionnaire," *BMC Cancer*, vol. 9, no. 1, p. 222, 2009.

[14] M. Pandey, B. C. Thomas, P. SreeRekha, K. Ramdas, K. Ratheesan, S. Parameswaran, B. S. Mathew, and B. Rajan, "Quality of Life determinants in women with breast cancer undergoing treatment with curative intent," *World J. Surg. Oncol.*, vol. 3, no. 1, p. 63, Sep. 2005.

[15] A. W. Awadalla, J. U. Ohaeri, A. Gholoum, A. O. Khalid, H. M. Hamad, and A. Jacob, "Factors associated with quality of life of outpatients with breast cancer and gynecologic cancers and their family caregivers: a controlled study," *BMC Cancer*, vol. 7, no. 1, p. 102, Jun. 2007.

[16] F. J. Alzabaidey, "Quality of Life Assessment for Patients with Breast Cancer Receiving Adjuvant Therapy," *J. Cancer Sci. Ther.*, vol. 04, no. 03, 2012.

[17] J. E. Bower, P. A. Ganz, K. A. Desmond, J. H. Rowland, B. E. Meyerowitz, and T. R. Belin, "Fatigue in Breast Cancer Survivors: Occurrence, Correlates, and Impact on Quality of Life," *J. Clin. Oncol.*, vol. 18, no. 4, pp. 743–743, Feb. 2000.
[18] J. Pallant, *SPSS Survival Manual: A step by step guide to data analysis using SPSS, 4th Edition*, 4 edition. Open University Press, 2010.
[19] B. G. Tabachnick and L. S. Fidell, *Using Multivariate Statistics*. HarperCollins College Publishers, 1996.
[20] M. Linting, J. J. Meulman, P. J. F. Groenen, and A. J. van der Koojj, "Nonlinear principal components analysis: introduction and application," *Psychol. Methods*, vol. 12, no. 3, pp. 336–358, Sep. 2007.
[21] A. B. Costello and J. W. Osborne, "Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis," *Pract. Assess. Res. Eval.*, vol. 10, pp. 173–178, 2005.
[22] B. G. Tabachnick and L. S. Fidell, *Using multivariate statistics*. Allyn and Bacon, 2001.

Appendices

Appendix 1: Correlation matrix for all the variables

	Q1	Q2	Q3	Q4	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28
Q1	1.00	0.60	0.40	0.39	0.60	0.32	0.20	0.35	0.33	0.28	0.18	0.09	0.15	0.15	0.06	0.00	0.13	0.42	0.05	0.09	0.07	0.13	0.27	0.10	0.01	0.22	0.13
Q2	0.60	1.00	0.58	0.54	0.65	0.42	0.49	0.34	0.44	0.25	0.33	0.14	0.18	0.18	0.05	-0.01	0.16	0.27	0.04	0.13	0.11	0.17	0.22	0.28	-0.01	0.07	0.12
Q3	0.40	0.58	1.00	0.31	0.56	0.30	0.05	0.59	0.27	0.42	0.30	0.21	0.25	0.24	0.14	-0.02	0.03	0.45	0.02	0.08	0.02	-0.01	0.18	0.17	0.00	0.14	0.04
Q4	0.39	0.54	0.31	1.00	0.48	0.52	0.19	0.20	0.53	0.22	0.08	-0.03	-0.05	-0.07	-0.03	0.27	0.19	0.23	0.32	0.25	0.17	0.05	0.17	0.24	0.27	0.28	0.10
Q6	0.60	0.65	0.56	0.48	1.00	0.50	0.13	0.62	0.46	0.34	0.24	0.09	0.05	0.06	0.14	-0.04	0.22	0.57	0.15	0.13	0.16	0.12	0.25	0.34	-0.03	0.15	0.13
Q7	0.32	0.42	0.30	0.52	0.50	1.00	0.06	0.15	0.61	0.17	0.25	0.14	0.02	0.01	0.03	-0.06	0.34	0.15	0.24	0.21	0.23	0.08	0.25	0.25	-0.06	0.13	0.10
Q8	0.20	0.49	0.05	0.19	0.13	0.06	1.00	0.00	0.17	-0.01	0.07	0.01	0.04	0.04	0.06	-0.02	0.08	-0.01	0.05	0.01	-0.05	0.23	0.09	0.13	-0.02	0.02	0.17
Q9	0.35	0.34	0.59	0.20	0.62	0.15	0.00	1.00	0.27	0.15	0.04	-0.04	-0.03	-0.03	0.18	0.03	0.11	0.69	0.01	0.14	0.04	-0.02	0.24	0.19	0.04	0.16	0.09
Q10	0.33	0.44	0.27	0.53	0.46	0.61	0.17	0.27	1.00	0.22	0.38	0.12	0.03	0.03	0.11	0.13	0.44	0.29	0.28	0.16	0.14	-0.01	0.20	0.18	0.13	0.23	0.14
Q11	0.28	0.25	0.42	0.22	0.34	0.17	-0.01	0.15	0.22	1.00	0.27	0.24	0.26	0.26	0.10	-0.02	0.03	0.22	0.35	0.07	0.04	0.09	0.03	0.21	-0.02	0.04	-0.03
Q12	0.18	0.33	0.30	0.08	0.24	0.25	0.07	0.04	0.38	0.27	1.00	0.44	0.44	0.43	0.35	0.02	0.43	0.09	0.07	0.09	0.00	-0.02	0.01	0.23	0.02	0.11	0.05
Q13	0.09	0.14	0.21	-0.03	0.09	0.14	0.01	-0.04	0.12	0.24	0.44	1.00	0.46	0.42	-0.03	-0.03	0.00	-0.06	-0.01	-0.01	-0.09	-0.04	-0.08	-0.03	-0.03	-0.04	-0.02
Q14	0.15	0.18	0.25	-0.05	0.05	0.02	0.04	-0.03	0.03	0.26	0.44	0.46	1.00	0.99	-0.04	-0.02	0.00	-0.06	-0.04	-0.04	-0.08	-0.03	-0.05	-0.05	-0.02	-0.01	0.01
Q15	0.15	0.18	0.24	-0.07	0.06	0.01	0.04	-0.03	0.03	0.26	0.43	0.42	0.99	1.00	-0.03	-0.01	0.00	-0.06	-0.04	-0.05	-0.07	-0.04	-0.06	-0.06	-0.01	-0.01	0.03
Q16	0.06	0.05	0.14	-0.03	0.14	0.03	0.06	0.18	0.11	0.10	0.35	-0.03	-0.04	-0.03	1.00	-0.02	0.23	0.20	0.01	0.04	0.00	0.07	0.03	0.24	0.02	0.37	0.09
Q17	0.00	-0.01	-0.02	0.27	-0.04	-0.06	-0.02	0.03	0.13	-0.02	0.02	-0.03	-0.02	-0.01	-0.02	1.00	0.07	0.33	0.12	0.33	0.00	0.08	0.03	0.00	1.00	0.39	0.08
Q18	0.13	0.16	0.03	0.19	0.22	0.34	0.08	0.11	0.44	0.03	0.43	0.00	0.00	0.00	0.23	0.07	1.00	0.20	0.11	0.09	0.05	0.10	0.09	0.12	0.07	0.16	0.15
Q19	0.42	0.27	0.45	0.23	0.57	0.15	-0.01	0.69	0.29	0.22	0.09	-0.06	-0.06	-0.06	0.20	0.33	0.20	1.00	0.05	0.25	0.09	0.06	0.25	0.14	0.35	0.47	0.17
Q20	0.05	0.04	0.02	0.32	0.15	0.24	0.05	0.01	0.28	0.35	0.07	-0.01	-0.04	-0.04	0.01	0.12	0.11	0.05	1.00	0.24	0.18	0.18	0.15	0.48	0.12	0.18	0.00
Q21	0.09	0.13	0.08	0.25	0.13	0.21	0.01	0.14	0.16	0.07	0.09	-0.01	-0.04	-0.05	0.04	0.33	0.09	0.25	0.24	1.00	0.62	0.35	0.55	0.20	0.33	0.26	0.11
Q22	0.07	0.11	0.02	0.17	0.16	0.23	-0.05	0.04	0.14	0.04	0.00	-0.09	-0.08	-0.07	0.00	0.00	0.05	0.09	0.18	0.62	1.00	0.19	0.48	0.12	0.00	0.18	0.12
Q23	0.13	0.17	-0.01	0.05	0.12	0.08	0.23	-0.02	-0.01	0.09	-0.02	-0.04	-0.03	-0.04	0.07	0.08	0.10	0.06	0.18	0.35	0.19	1.00	0.24	0.30	0.08	-0.01	0.13
Q24	0.27	0.22	0.18	0.17	0.25	0.25	0.09	0.24	0.20	0.03	0.01	-0.08	-0.05	-0.06	0.03	0.03	0.09	0.25	0.15	0.55	0.48	0.24	1.00	0.31	0.03	0.14	0.08
Q25	0.10	0.28	0.17	0.24	0.34	0.25	0.13	0.19	0.18	0.21	0.23	-0.03	-0.05	-0.06	0.24	0.00	0.12	0.14	0.48	0.20	0.12	0.30	0.31	1.00	0.01	0.18	0.15
Q26	0.01	-0.01	0.00	0.27	-0.03	-0.06	-0.02	0.04	0.13	-0.02	0.02	-0.03	-0.02	-0.01	0.02	1.00	0.07	0.35	0.12	0.33	0.00	0.08	0.03	0.01	1.00	0.42	0.08
Q27	0.22	0.07	0.14	0.28	0.15	0.13	0.02	0.16	0.23	0.04	0.11	-0.04	-0.01	-0.01	0.37	0.39	0.16	0.47	0.18	0.26	0.18	-0.01	0.14	0.18	0.42	1.00	0.18
Q28	0.13	0.12	0.04	0.10	0.13	0.10	0.17	0.09	0.14	-0.03	0.05	-0.02	0.01	0.03	0.09	0.08	0.15	0.17	0.00	0.11	0.12	0.13	0.08	0.15	0.08	0.18	1.00

Appendix 2: Correlation matrix for the remaining variables

	Trouble doing strenuous activity	Trouble taking long walk	Trouble taking short walk	Stay in bed /chair during the day	Limited in doing daily activities	Limited in pursuing your hobbies	Pain interfere with daily activities	pain	Need rest	Felt weak	Trouble sleeping
Trouble doing strenuous activity	1.00	0.58	0.38	0.44	0.60	0.38	-0.50	-0.33	-0.35	-0.15	-0.32
Trouble taking long walk	0.58	1.00	0.55	0.58	0.67	0.47	-0.38	-0.35	-0.47	-0.42	-0.29
Trouble taking short walk	0.38	0.55	1.00	0.36	0.53	0.34	-0.40	-0.55	-0.30	-0.26	-0.43
Stay in bed /chair during the day	0.44	0.58	0.36	1.00	0.51	0.52	-0.40	-0.24	-0.52	-0.31	-0.30
Limited in doing daily activities	0.60	0.67	0.53	0.51	1.00	0.62	-0.59	-0.53	-0.52	-0.36	-0.41
Limited in pursuing your hobbies	0.38	0.47	0.34	0.52	0.62	1.00	-0.32	-0.20	-0.56	-0.38	-0.23
Pain interfere with daily activities	-0.50	-0.38	-0.40	-0.40	-0.59	-0.32	1.00	0.64	0.35	0.22	0.35
Pain	-0.33	-0.35	-0.55	-0.24	-0.53	-0.20	0.64	1.00	0.31	0.19	0.29
Need rest	-0.35	-0.47	-0.30	-0.52	-0.52	-0.56	0.35	0.31	1.00	0.59	0.27
Felt weak	-0.15	-0.42	-0.26	-0.31	-0.36	-0.38	0.22	0.19	0.59	1.00	0.18
Trouble sleeping	-0.32	-0.29	-0.43	-0.30	-0.41	-0.23	0.35	0.29	0.27	0.18	1.00

Appendix 3: Scree plot

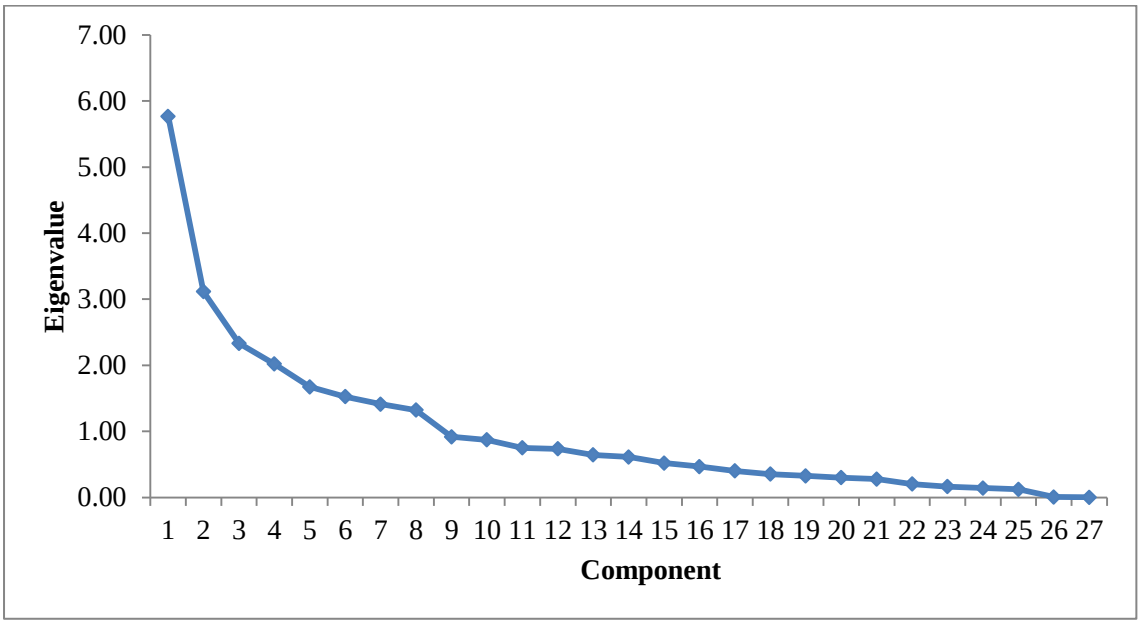

Figure 7: Initial scree plot

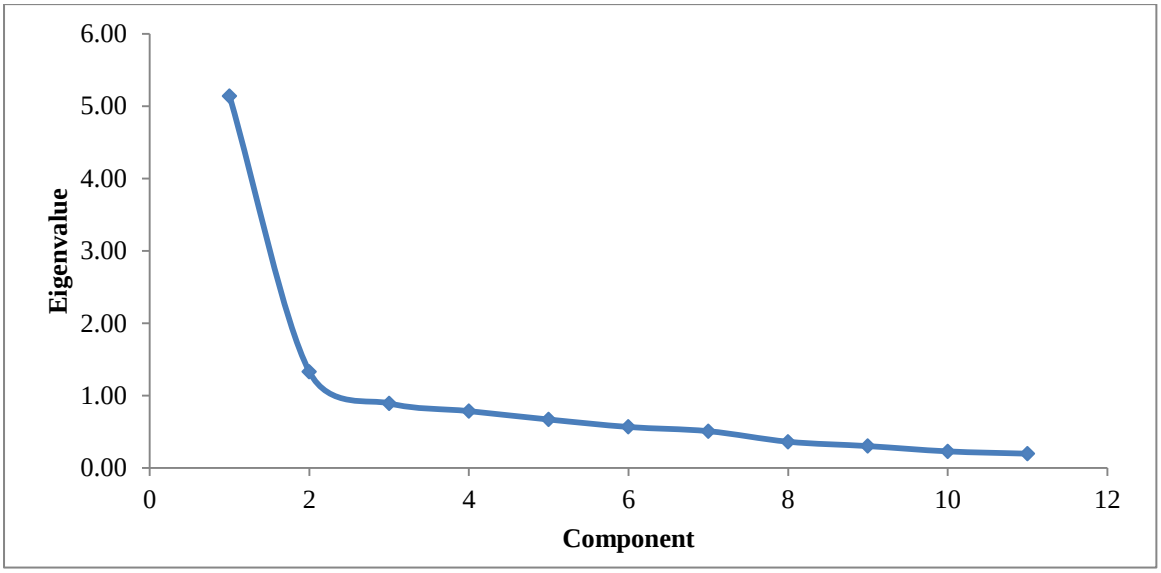

Figure 8: Scree plot after some variables are dropped