University of Fort Hare

Together in Excellence

Statistical methods to model the influence of age and gender on the behavioral risk factors of HIV/AIDS

A half dissertation submitted in partial fulfillment of the requirements of the degree of

Master of Science

in

Biostatistics and Epidemiology

At the University of Fort Hare, Department of Statistics.

By

Boikhutso Tlou

November 2010

Declaration

I hereby declare that the content of this research work is my original work. Information extracted from other sources is acknowledged as such. To this end all the resources used have been duly acknowledged.

Signature……..

Date 07 November 2010

Acknowledgement

To begin with, I would like to give thanks unto the almighty God, for giving me strength, wisdom and guidance throughout this dissertation.

My greatest appreciation goes to my supervisor Professor Y.Qin, for his support and guidance throughout the course of this dissertation. I would also like to thank him for the assistance, knowledge, experience and courage he gave me throughout this dissertation.

I want also to thank Professor J.C. Tyler, for her support in providing me with the needed material for this study.

To my lovely daughter Tebogo, I would like to say you are a superstar. You inspired me to work hard and I dedicate this research to you. Also, not forgetting my close friend and mother of my baby Pamellar Pozisa Mbongonya for her moral support, “Enkosi Mamkhwananzi”.

Furthermore, I would like to thank my family for the support and perseverance they gave me throughout my studies. I would like to say this to them, “Kealeboga ba ga TLOU”.

To my friends, colleagues and the Statistics Department, I would like to say thank you.

Abstract

The effects of gender and age on the behavioral risk of HIV/AIDS are not clearly understood as previous distinct studies which have been carried out, have given disputable and contradictory outcomes. This study therefore, discusses the statistical methods which can be used to model the influence of age and gender on the behavioral risk factors of HIV/AIDS. In general, generalized linear models are the main methods which can be applied to depict the impact of age and gender on the behavioral risk of becoming infected with HIV/AIDS virus. In this study, the main methods used were logistic regression, log-linear regression and multiple regressions. Behavioral risk was taken as the dependent variable while age, gender, number of sexual partners, religious beliefs and alcohol and drug abuse were fitted as predictor variables. The three statistical methods gave significant results for gender and insignificant results for age. Furthermore, comparisons were made on the three regression methods and the logistic regression gave the best results. It was therefore concluded that gender plays a significant role on the behavioral risk of HIV/AIDS. The results of the study showed that gender of the student and number of sexual partners had a significant effect on the risk behavior of the university students. In future, it may be very important to find out why age is not a significant factor on risk behavior of HIV/AIDS among university students.

Table of Contents

CHAPTER ONE

Introduction

This chapter will introduce the study and the main problem statements of the study. It will also describe the importance and significance of the study.

HIV/AIDS prevalence in Africa

The risk of the Human Immunodeficiency Virus (HIV) and Acquired immunodeficiency Syndrome (AIDS) has been a major global problem. Even though HIV/AIDS' prevalence is stabilizing and declining in some countries such as Malawi, Southern Africa still remains the epicenter of the deadly disease with about 35% of new infections and 38% of HIV/AIDS deaths[1]. Young women are more prone to the fatal disease as they are three times more likely to be HIV positive than young men.

Children in Africa comprise approximately 90% of the estimated 2 million children under the age of 15 living with HIV/AIDS globally [1].
South Africa on its own was estimated to have 5.0 million people living with HIV/AIDS in 2009 more than any country in Southern Africa [31]. According to [32], it was believed that in 2008 more than 250,000 South Africans died of AIDS. Approximately one-in-three women aged 25-29, and over a quarter of men aged 30-34, are living with HIV [33].

HIV/AIDS is a major if not the principal factor in the overall rising number of deaths in Africa as a continent. HIV/AIDS still continues to spread rapidly; it now causes deaths more than any other infectious disease. It is now rated as the

fourth worst hazard in the world (after heart disease, stroke and respiratory diseases) and has become the largest cause of death in Africa [1].

1.2 Analysis of factors affecting HIV/AIDS

Firstly, gender inequality is one of the important factors that influence the spread of HIV/AIDS. Scholars and experts have devoted increasing attention to the role of gender equality in the spread of HIV/AIDS. It is believed that the marginal social location and low status of women in many societies explains their weakening ability to protect themselves from the virus. The sexual vulnerability of young women in makes them even more vulnerable to HIV/AIDS than their other counterparts

Secondly, poverty is associated with low endowments of human and financial resources, such as low levels of education with associated low levels of literacy and few marketable skills, generally poor health status and low labour productivity as a result. The poor health status amongst many Africans is now globally recognized as a very significant co-factor in the transmission of HIV/AIDS.

It is not surprising that in these circumstances some poor people may adopt behaviors which expose them to HIV/AIDS infection. In most African countries the practice of prostitution is a source of income, whilst on the other hand this practice exacerbates the spread of HIV/AIDS.

Thirdly, Africa as a continent has diverse cultural beliefs and practices, which influence HIV/AIDS prevention, education, and transmission across the continent. In some countries such as South Africa, Zimbabwe and Swaziland, polygamy is still widely practiced and tends to be an obstacle in trying to prevent HIV/AIDS. It is difficult to practice the one partner theory in such countries.

Moreover, there are biological factors which influence the spread of HIV/AIDS. Sexual transmission of HIV normally occurs through the dendritic cells in the male and female genital tract. Such cells are in abundance in the transformation zone of the cervix in women, and penile urethra in men. This abundance suggests that genital tract infections would increase the sexual transmission of HIV/AIDS.

It is also important to highlight the fact that youths, including students in tertiary institutions, are the most affected by the HIV/AIDS epidemic because of indulging in unprotected sex and drug abuse .As a result, this study will be carried out among university students. There are quite a number of behavioral risk factors of HIV such as

- Not taking antiretroviral drugs properly
- Mixing sex and drugs/alcohol
- Unsafe drug use behavior
- Unsafe sexual behavior

These risk behaviors have been the main factors leading to HIV [2].

Also, it is worth mentioning that even though the main predictor variables will be age and gender on the response variable behavioral risk, some other explanatory variables will be included to monitor the interactional effects. The complete list of explanatory factors which will be used in this research comprises:

- Gender
- Age
- Alcohol and Drug abuse
- Number of partners
- Religious beliefs

However, our main emphasis on this study will be on the influence of gender and age on the behavioral risk of HIV/AIDS.

1.3 Motivation and importance

In this thesis, we will be discussing the statistical methods that can be used to show the influence of gender and age on the behavioral risk factors of HIV/AIDS. In simpler terms, this research is about finding suitable statistical methods that can be used to model the impact of gender and age on HIV/AIDS behavioral risks.

The roles of gender and age on the behavioral risk of HIV/AIDS are not clearly understood as several previous studies have given disputable and contradictory outcomes.

As a result, there is a demonstrable need to investigate, analyze and understand how previous studies differ so that at the end of this research, we can make explicit conclusions about the influence of gender and age on the behavioral risk factors of HIV/AIDS.

It is important to mention the fact that this research will be carried among university students. We targeted university students mainly because, students might be viewed as the future of the African continent as a whole and their high literacy levels might assist in reducing the spread the message of HIV/AIDS.

Medical, social and educational institutions have been alerting the public about the hazardous impacts of the various risk behaviors. Thus, this study will explore the influence of gender and age on risk behaviors towards HIV/AIDS.

Moreover, the prevalence of HIV/AIDS in Africa is still escalating at an alarming rate [3]. More still, this research will go a long way in trying to link the previous statistical studies done with the recent ones on the influence of gender and age on behavioral risk factors of HIV/AIDS.

It is important to give a brief description of the statistical methods that will be applied in this thesis.

1.4. Statistical methods that will be applied

The variables gender and age will be taken as predictor variables, though other variables will be included to investigate the interaction effects. Our response variable will be taken as behavioral risk, which might be categorical or continuous.

There are many statistical methods that can be used to model a categorical or continuous response variable against categorical or continuous explanatory variables.

These statistical methods include the following:

- Logistic regression – This model is a type of an explanatory or predictive model for a response variable that is categorical with just two categories.
- Loglinear regression - If the explanatory and response variables are categorical data then one can use loglinear regression to model data Loglinear models are mostly used when a model describes and explains association patterns among a set of categorical response variables.

In short we can say when there is no distinction between the response and explanatory variable, the loglinear model provides a good statistical analysis for testing associations and interactions among set of categorical response variable [9].

- Multiple regressions - It gives the relationship between several independent or explanatory variables and a dependent or response

variable. In general it allows the simultaneous testing and modeling of multiple independent variables.

- Analysis of Variance (ANOVA) – The statistical tests carried out in ANOVA are based on the F-ratio (the variation due to an experimental treatment or effect divided by the variation due to experimental error.

- Factor Analysis – is a correlation technique to determine meaningful clusters of shared variances.

The statistical methods will be discussed in detail in chapter three.

Aims and Objectives

The main objective of this thesis is to explain the statistical methods that can be used to show the influence of gender and age on the behavioral risk factors of HIV/AIDS.

This research will make comparisons of different statistical methods in depicting the influence of gender and age on risk behavior of HIV/AIDS. At the end of the study we need to choose the suitable method for modeling the influence of age and gender on risk behavior of HIV/AIDS.

The aims and objectives of the study can therefore be summarized as follows:

- ❖ To fit a hierarchical log-linear model on behavioral risk in order to assess the statistical patterns of association among variables.
- ❖ To fit a logistic regression of a binary categorical response variable to a set of categorical explanatory variables.
- ❖ To fit multiple regressions to the response variable behavioral risk

- ❖ Making comparisons on the impact of age and gender on behavioral risk using the above named statistical methods.
- ❖ Monitoring the effect of gender roles or societal expectations, personal gender ideologies and gender based power differentials

1.6 Structure of the thesis

We will briefly outline what will be contained in the preceding chapters.
In chapter two the literature review of previous and existing research on the influence of age and gender on behavioral risk is discussed.

In chapter three the methods of the research will be outlined. In chapter four of the research the results of statistical analyses are presented. In chapter five conclusions on the findings of the study will be made.

We will give the definitions of the key terms of the research. Gender can be defined as the set of ideas shared by people belonging to a given group or population [4].In this research we need to see whether the set of ideas possessed by people of different sex have an impact on the risk behavior of HIV/AIDS.

CHAPTER TWO

Literature review

This chapter will give the relevant statistical literature review related to the study. I Previous studies related to the study will be reviewed. A brief discussion and description of the literature of generalized linear models will be presented in this chapter. However, detailed application of the statistical methods used will be presented in chapter three.

2.1 Introduction

The risk of HIV/AIDS has been a major global problem. Various studies have been conducted to illustrate the influence of gender and age on the behavioral risk factors of HIV. In this chapter we discuss the statistical methods that have been used to show the effect of gender and age on the behavioral risk factors of HIV/AIDS.

This chapter will review the statistical literature for the influence of the above mentioned variables on the behavioral risk factors of HIV/AIDS.

2.2 Generalized Linear models (GLM)

Generalized Linear Models include special cases such as linear regression, analysis of variance, logit and probit models, loglinear models and multinomial response models [8].

The general linear model is a generalization of linear models. A linear model specifies the relationship between a dependent (or response) variable Y, and a set of predictor variables, the X's, so that we calculate fitted values

$$Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_n X_n$$

In the equation b_0 is the regression coefficient for the intercept and the b_i values are the regression coefficients for variables 1 through n computed from the data. As an example we could predict HIV/AIDS as a function of gender and age, where gender is a dummy variable.

According to Nelder [6], there are three components which specify a generalized linear model namely:

- Random component - it identifies the probability distribution of the response variable.
- Systematic component - it specifies a linear function of explanatory variables that is used as a predictor.
- Link function - it describes the functional relationship between the systematic component and the expected value of the random component.

The random component consists of the response variable Y with n independent observations ($y_1, y_2, y_3 ..... y_N$) from a distribution in the exponential family with the probability density function $f(y_i, \theta_i) = a(\theta_i) b(y_i) e^{y_i Q(\theta_i)}$, where $Q(\theta_i)$ is the natural parameter of the distribution and θ_i is a parameter which may vary for i = 1, 2 …N depending on the explanatory variables.

There are some distributions like the Binomial and Poison which are part of this family. According to Agresti [7] the link function connects the random and systematic components. A case in point we assume that $\mu_i = E(Y_i)$ and $g(.)$ be a differentiable link function. This function will relate $E(Y_i)$ to the explanatory variables through the following expression [8] $\eta_i = g(\mu_i) = \sum_j \beta_j x_{ij}, \quad i = 1, 2, ..., N$

The first generalized linear model to review will be the logistic regression.

2.2.1 Logistic regression

Logistic regression describes the relationship between a categorical response and a set of predictor variables. A categorical response can be a binary variable, nominal variable or ordinal variable. Each type of categorical variables requires different techniques to model its relationship with the predictor variables.

Logistic regression is a form of statistically modeling that is often appropriate for categorical outcome variables. It may be described as an optimal method for the regression analysis of dichotomous (binary) dependent variables. In this study, logistic regression will be examined when the response variable is dichotomous. The central mathematical concept that underlies logistic regression is the logit- the natural logarithm of an odds ratio.

To fully appreciate the logistic regression, it is very important to have a clear understanding of odds and odds ratios. The odds of an event are the ratio of the expected number of times that an event will occur to the expected number of times it will not occur. An odds of 5 means we expect 5 times as many occurrences as non-occurrences. There is a simple relationship between probabilities and odds. If p is the probability of an event and O is the odds of the event, then

$$O = \frac{p}{1-p} = \frac{probabilityofevent}{probability\ of\ no\ event} \qquad p = \frac{O}{1+O}$$

In table 1, odds less than 1 correspond to probabilities below 0.5, while odds greater than 1 correspond to probabilities greater than 0.5.In parallel to probabilities, odds have a lower bound of 0, but unlike probabilities, there is no upper bound on the odds. Odds provide a more sensible scale for multiplicative comparisons.
The functional relationship is illustrated in table 1:

Table 1 Relationship between odds and probability

Probability	Odds
0.1	0.11
0.2	0.25
0.3	0.43
0.4	0.67
0.5	1.00
0.6	1.50
0.7	2.33
0.8	4.00
0.9	9.00

An example is a case where by the probability of Tebogo passing her masters is 0.3 and the probability of John passing is 0.6.We can claim that John's probability is twice as great as that of Tebogo. But if Tebogo's probability is 0.6, it's impossible for John's probability to be twice as great. However, there is no problem on the odds scale. A probability of 0.6 corresponds to odds of $\frac{0.6}{1-0.6}=1.5.$ Doubling that yields odds of 3.Converting back to probabilities gives us $\frac{3}{1+3}=0.75$ .Odds ratios, are a widely used measure of the relationship between dichotomous categorical response variable and one or more categorical or continuous predictor variables. The corresponding Table 2 explains odds ratio

Table 2 Sample data for Gender and Recommendation for Remedial Reading Instruction

	Gender		
Remedial reading instruction	Boys	Girls	Total
Recommended (coded as 1)	73	15	88
Not recommended (coded as 0)	23	11	34
Total	96	26	122

If we assess the boy's odds of being recommended for remedial reading instruction relative to a girl's odds, the result is an odds ratio of 2.33, which suggests that boys are 2.33 times more likely, than not to be recommended for remedial reading classes compared with girls. The odds ratio is derived from two odds (73/23 for boys and 15/11 for girls).

2.2.2 The Logit model

One of the major problems with the linear probability model is that probabilities are bounded by 0 and 1, but linear functions are inherently unbounded. The solution is to transform the probability so that it's no longer bounded. When we are transforming the probability to odds we remove the upper bound. If the logarithm of the odds is taken, the lower bound is also removed. Setting the result to a linear function of the explanatory variables, we get the logit model. For k explanatory variables and $i=1,...n$ individuals, the model is

$$\log\left[\frac{p_i}{1-p_i}\right]=\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+..+\beta_k x_{ik}$$

Where p_i is, the probability that $y_i=1.$ The expression on the left-hand side is usually referred to as the logit or log-odds. The logistic regression is well suited for describing and testing hypotheses about relationships between a categorical response variable and one or more categorical or continuous predictor variables. The simple logistic model has the form

$$\log it(Y)=\ln\left(\frac{\pi}{1-\pi}\right)=\alpha+\beta x. \qquad (1)$$

Taking the antilog of (1) on both sides, one derives an equation to predict the probability of the occurrence of the outcome of interest as follows:

$$\pi=\Pr obability(Y=outcome\,needed \mid X=x, a\,specific\,value\,of\ X)=\frac{e^{\alpha+\beta_x}}{1+e^{\alpha+\beta_x}} \qquad (2)$$

where π is the probability of the outcome of interest or event α is the Y intercept and β is the regression coefficient. X can be continuous or categorical but Y is always categorical.

Considering n independent Bernoulli random variables $Y_1,.....Y_n$ having observed values $y_0 = (y_{01},....y_{0n})'$.

Each observation $i = 1,.....n,\ x_i = (x_{i1},...., x_{ip}, x_{i,p+1},..., x_{i,p+q})'$ be a vector of p + q explanatory variables, and denote $X = (x_1,...., x_n)'$. $\pi_i = \pi(x_i) = \Pr(Y_i = 1/ x_i)$ be the event probability for each i = 1,…..n and denote $\pi = (\pi_1,..., \pi_n)'$. $E(y_0) = \pi_0$

Then the logistic regression model is $g(\mu) = g(\pi)$ = logit $(\pi) = X\beta$, or

logit $(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = x'_i\,\beta$, where $\beta = (\beta_1,.....\beta_{p+q})'$ is the unknown parameter vector. The joint probability of the observed y_0 is a product of n Bernoulli functions:

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{\,y_{0,i}} (1-\pi_i)^{1-y_{0,i}}$$

Because $\pi_i = \frac{e^{x_i\beta}}{1+e^{x_i\beta}}$ we obtain $L(\beta) = \frac{\exp(y'_0\, X\beta)}{\prod_{i=1}^{n}[1+\exp(x_i\beta)]}$

In our own study the categorical response variables can be taken as the behavioral risk factors of HIV/AIDS. These include the following:

- Sexual behavior
- Drug use behaviors
- Combinations of sex and drug/alcohol behaviors

There are several predictor variables which can be employed in this study, but the main emphasis will be on the influence of gender and age though the fitting will be done with many other explanatory variables.

A large positive regression coefficient means that the risk factor strongly influences the probability of that outcome; while a near zero regression coefficient means that the risk factor has little influence on the probability of that outcome [11]. The goal of the logistic regression is to correctly predict the

category of outcome using the most parsimonious model. There are many distinct available options which can be used during model selection.

Variables can be entered into the model in the order specified by the researcher or alternatively logistic regression can test the fit of the model after each coefficient is added or deleted, called stepwise logistic regression [12].

We can only use stepwise logistic regression for exploratory analysis and not for theory testing [13]. Theory testing can be defined as the testing of hypotheses of the relationships between variables. The most preferred method for exploratory analysis is backward stepwise regression, in which the analysis starts with a saturated model and variables are eliminated from the full model in an iterative process. The suitability of the model is checked after the elimination of each variable to test for goodness, this process will be explained in detail in chapter three [14].

The two main functions of the logistic regression are:

- Prediction of group membership- Since we know that the logistic regression calculates the probability of success over the probability of failure, the results of the analysis will be in the form of odds ratio. [15]
- Descriptions of the relationships and strengths among the variables (for example drinking 20 beers a day puts a person at a higher risk for developing liver cirrhosis than risk association with smoking)

2.2.3 Advantages of using logistic regression over ordinary linear regression

- If we use linear regression, the predicted probability values will become greater than 1 and less than zero when X moves far enough on any X-axis.

- ❖ We know that one of the assumptions of regression is that variance of Y is constant across the values of X. However, this condition cannot hold with binary variables since variance is PQ, where Q = 1-P [16].
- ❖ The significance testing rests upon the assumption that errors of prediction are normally distributed. Because Y only takes values 1 and 0, this assumption is not justified.

Next, we discuss the loglinear regression as the second generalized linear model to be used in this study.

2.3 Loglinear Models

Loglinear models are another important tool for the analysis of categorical data. Loglinear model methodology is appropriate when there is no clear distinction between response and explanatory variables, for example when all of the variables are observed simultaneously. The loglinear model point of view treats all variables as response variables, and the focus is on statistical independence and independence. Loglinear modeling of categorical data is analogous to correlation analysis for normally distributed response variables and is useful in assessing patterns of statistical dependence among subsets of variables.

A family of log-linear models is often refereed to as an exponentially family. If we consider n independent, identically distributed discrete random variables $X_1,.....,X_n$ with a common point probability:

$$f(x \mid \theta_1,...\theta_k) = P(X_i = x \mid \theta_1,....., \theta_k), \qquad (1)$$

which depends on k real valued parameters $\theta_1,.....\theta_k$. The model:

$$f(x_1,...x_n \mid \theta_1,...\theta_k) = \prod_{i=1}^{n} f(x_i \mid \theta_1,...., \theta_k) \qquad (2)$$

is then called a log-linear model or is said to form an exponential family, if the logarithm of (1) has the functional form :

$$\ln f(x|\theta_1,....\theta_k) = \sum_{j=1}^{m} g_j(x)\varphi_j(\theta_1,....\theta_k) + h(x) - K(\theta_1,....,\theta_k) \qquad (3)$$

where $g_j, \varphi_j, and\ h$ are all real valued functions of their arguments. The function K satisfies:

$$K(\theta_1,...\theta_k) = \ln\{\sum_x \exp(\sum_j g_j(x)\varphi_j(\theta_1,.....,\theta_k) + h(x))\} \qquad (4)$$

since $\sum_x f(x|\theta_1,....\theta_k) = 1.$

The dimension of the exponential family is the smallest integer m for which the representation (3) is possible. The dimension of an exponential family is less than the apparent dimension of the logarithmic form (3) if there are linear dependencies between either the $g's or\ the\ \varphi's.$

Although the simplest application of loglinear models is in testing statistical independence between two categorical values, the methodology is most useful in situations in which there are several variables. In this chapter a brief discussion of two way contingency tables and hierarchical approach to loglinear modeling will be discussed.

2.3.1 Loglinear modeling for the $s \times r$ Tables

If a sample of n observations is categorized to two categorical variables, one having s levels and the other having r levels, the resulting frequencies can be displayed in an $s \times r$ contingency table:

Table 3 $s \times r$ contingency table

Level of X	*Level of Y*				**Total**
	1	2	…	r	
1	n_{11}	n_{12}	…	n_{1r}	n_{1+}
2	n_{21}	n_{22}	…	n_{2r}	n_{2+}
.	.	.	…		
.	.	.	…		
.	.	.	…		
s	n_{s1}	n_{s2}	…	n_{sr}	n_{s+}
Total	n_{+1}	n_{+2}	…	n_{+r}	n

The derivation of the loglinear model from the 2×2 table to the $s\times r$ table is clear and straight forward when using Table 1.

The saturated model for a 2×2 table is

$$\log(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \quad i = 1,....,s,\ j = 1,...,r$$

where $m_{ij} = n\pi_{ij}$ is the expected frequency in the (i, j) cell. The parameter μ is fixed by the sample size n and the model has $s + r + sr$ parameters $\lambda_i^X, \lambda_j^Y, and\ \lambda_{ij}^{XY}$.

The sum to zero constraints

$$\sum_{i=1}^{s}\lambda_i^X = 0 \qquad \sum_{j=1}^{r}\lambda_j^Y = 0 \qquad \sum_{i=1}^{s}\lambda_{ij}^{XY} = \sum_{j=1}^{r}\lambda_{ij}^{XY} = 0$$

implies $(s-1)+(r-1)+(s-1)(r-1) = sr-1$ parameters and zero df for testing lack of fit. Denoting $\hat{m}_{ij} = n_{i+}n_{+j}/n,$ the likelihood ratio statistic

$$G^2 = 2\sum_{i=1}^{s}\sum_{j=1}^{r} n_{ij}\log(n_{ij}/\hat{m}_{ij})$$

tests the null hypothesis $H_0 : \lambda_{ij}^{XY} = 0,\ for\ i = 1,.....,s-1,\ j = 1,...,r-1.$ Under the null hypothesis of independence, G^2 has an approximate chi-square distribution with $(s-1)(r-1)\ df$.

If H_0 is true, the reduced model $\log(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y$ is the model of independence of $X\ and\ Y$. This model has $(s-1)+(r-1)$ linearly independent λ parameters and $(s-1)(r-1)\,df$ for testing lack of fit.

Loglinear models can be used to analyze the relationship between two categorical variables (two-way contingency tables), though they are more commonly used to evaluate multiway contingency tables that involve three or more categorical variables. The variables investigated by loglinear models are all treated as response variables. In other words there is no distinction made between response and explanatory variables. Therefore loglinear models only demonstrate association between variables.

Supposed we are interested in the relationship between sexual behavior, gender and age, we could take a sample of our subjects and for each subject determine the gender, determine sexual behavior as protected sex or unprotected sex, and approximate age into some categories.

Given a three dimensional contingency table the model assumes a sample size n distributed over $I\,J\,K = N$ cells. Under sampling (no fixed margins), the probability that an observation (X,Y,Z) will fall into cell $i\,j\,k$ is then π_{ijk} for all $i = 1,.....I,\ j = 1,.....J, k = 1,....., K.$

The expected value m_{ijk} thus is $n\pi_{ijk}$.

Mutual independence of the three variables will be now equivalent to $\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$ as well as $m_{ijk} = n\pi_{i++}\pi_{+j+}\pi_{++k}$,where π_{++} is the marginal probability of category i on variable X ,Then

$$\log(m_{ijk}) = \log(n) + \log(\pi_{i++}) + \log(\pi_{+j+}) + \log(\pi_{++k})$$

which is equivalent to

$$\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z \qquad \text{(i)}$$

with

$$\lambda_i^X = \log(\pi_{i++}) - \sum_v \log(\pi_{v++})/I$$

$$\lambda_j^Y = \log(\pi_{+j+}) - \sum_v \log(\pi_{+v+})/J$$

$$\lambda_k^Z = \log(\pi_{++k}) - \sum_v \log(\pi_{++v})/K$$

$$\mu = \log(n) + \sum_v \log(\pi_{v++})/I + \sum_v \log(\pi_{+v+})/J + \sum_v \log(\pi_{++v})/K.$$

The parameters $\{\{\lambda_i^X\},\{\lambda_j^Y\}$ *and* $\{\lambda_k^Z\}\}$ satisfy

$$\sum \lambda_i^X = \sum \lambda_i^Y = \sum \lambda_k^Z = 0 \qquad \text{(ii)}$$

Model (i) is the model of mutual independence for a three dimensional contingency table. Without the constraints (ii) it would not be possible to identify the parameters uniquely.

In parallel to the classical ANOVA modeling, interactions between two or all three variables can be modeled. Introducing the additional terms $\{\lambda_{ij}^{XY}\},\{\lambda_{ik}^{XZ}\},\{\lambda_{jk}^{YZ}\}$,

$$\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}.$$

2.3.2 The Hierarchical Approach to Loglinear Modeling

In order to show the hierarchical approach to loglinear modeling, we shall assume a 2 * 2 * 2 muIti-way contingency table consisting of three variables, (X,Y,Z) each with two levels. The model structure can be illustrated as shown below

$$Ln(z_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$$

A hierarchy of models exists whenever a complex multivariate relationship present in the data necessitates inclusion of less complex interrelationships [19].

For instance, in the above mentioned equation, if a three – way interaction is present (X,Y,Z), the equation for the model must also include all two way effects (XY, XZ, YZ) as well as the single variable effects (X, Y, Z) and the overall mean (μ).

Log linear analysis is different from Logistic regression in three different ways:

- The expected distribution of the categorical variables is Poisson, not binomial or multinomial.
- The link function is the natural log of the response, not the logit of the response variable as in logistic regression.
- Predictions are estimates of the cell counts in a contingency table, not the logit of the expected value of a Bernoulli variable

2.3.3 Statistical independence and association

In this research we want to see the effect of gender and age on the behavioral risk factors of HIV/AIDS .It is very important to see whether gender and age are depend on any given explanatory variable under study.

When two events are statistically independent, independence means that knowing whether one of the events occurs or not makes it neither more probable nor less probable that the other occurs. In other words, the occurrence of one event does not affect the probability of the occurrence of the other event.

Similarly, when we assert that two random variables are independent, we intuitively mean that having knowledge about the value of either of them does not yield any information about the value of the other [20].

There are several methods that are used to test for statistical independence of variables. In situations where both variables are response variables, we can describe the association using their joint distribution, the conditional distribution of Y given X, or the conditional distribution of X given Y. The relationship between the conditional distribution of Y given X and the joint distribution is as shown below:

$$\pi_{j/i} = \pi_{ij} / \pi_{i+}, \forall i, j .$$

The variables are statistically independent if all joint probabilities equal the product of the marginal probabilities, that is, if

$$\pi_{ij} = \pi_{i+}\pi_{+j} \quad , \forall i = 1, ..., I \; and \; j = 1, ..., J .$$

When X and Y are independent, $\pi_{j/i} = \pi_{ij} / \pi_{i+} = \pi_{+j}, \; for \; i = 1, ..., I,$ and $j = 1, ..., J.$ In this case each conditional distribution of Y is identical to the marginal distribution of Y. Thus, two variables are independent when the probability of column response j is the same in each row, for j = 1…, J [7].

2.3.4 The direction and the strength of the association

When dealing with association or dependence of variables, it is also important also measure the strength and direction of that association. We can use correlation coefficients to measure the strength of association between two variables [21]

Since, in our study we will be mainly dealing with response variables with binary responses, our main emphasis will be on the comparisons of the binary responses. There are three methods which can be used which are: relative risk, difference of proportions, and odds ratios.

2.3.5 Difference of proportions (DP)

Given a response Y and a predictor X we can compare two levels of X at a given level of Y using: DP = $\pi_{j/h} - \pi_{j/i}, \forall h \neq i \; levels \; of \; X$. The difference of proportions is in the range -1.0 to 1.0. However the difference of proportions is not the best measure for comparing proportions in all cases. In fact, for proportions close to zero and one, the DP is not a good measure; instead the ratio would be more appropriate.

2.3.6 Relative risk (RR)

Frequently in the statistical analysis of binary outcomes (X, Y) the outcome of interest has a relatively low probability. The relative risk (RR) is particularly attractive because it can be calculated by hand in the simple case, but is also amenable to regression modeling, typically in a Poisson regression framework.

For 2 x 2 contingency tables, the relative risk is the ratio RR= $\pi_{1/1} / \pi_{1/2}$. This ratio can be any non -negative real number. When using relative risk, a value of one implies independence of variables

2.3.7 Odds ratio (OR)

The odds ratio is a measure of effect size, describing the strength of association or dependence between two binary data values. It is used as a descriptive statistic and plays an important role in logistic regression. Unlike other measures of association for paired binary data, such as the relative risk, the odds ratio treats the two variables being compared symmetrically, and can be estimated using some types of non random samples.

Given a 2 x 2 contingency table, the odds ratio compares the odds of response 1 proportion for group 1 to the odds of response1 proportion for group 2. It can be calculated as illustrated below:

$$\text{OR}= \frac{\pi_{1/1} / \pi_{1/2}}{\pi_{2/1} / \pi_{2/2}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

The values ranges from zero to infinity, when OR = 1, it means there is no association between the row variable and the column variable, that is, X and Y are independent.

When 1< OR<∞, subjects in row 1 are more likely to make the first response than the subjects in row 2.
When 0<OR<1, it means that the first response is less likely in row 1 than in row 2.
Odds ratios can also be useful for describing contingency tables larger than 2 x 2. Odds ratios for I*J tables can use each of the $\binom{I}{2}$ pairs of rows in combination with each of the $\binom{J}{2}$ pairs of columns. A subset of (I-1) (J-1) local odds ratios determines all the $\binom{I}{2}\binom{J}{2}$ odds ratios formed from pairs of rows and pairs of columns.

As a result, the general OR for the I*J contingency tables can be written as follows:

$$OR_{ij} = \frac{\pi_{i,j}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}},\ i = 1,.....,I-1,\ j = 1,.....,J-1.$$

2.3.8 Relationship between OR and RR

The relationship between OR and RR can be described as:

$$RR = \frac{\pi_{11}}{\pi_{12}}.\frac{\pi_{12}+\pi_{22}}{\pi_{11}+\pi_{12}}$$
$$= \frac{\pi_{11}}{\pi_{12}}.\left(\frac{\pi_{12}}{1-\pi_{.2}}\right)$$

2.3.9 Special cases of loglinear models

In this section we will show that logit models, e.g., logistic regression, are a special case of log-linear models where there is a single response variable of interest; thus we can fit the loglinear model when we have a clear response, but account for that when we do the interpretation of the results. However, if there are more than two responses we will use log-linear models.

As a result, it is important to show the development of the loglinear model that moves away from the joint response status of the variables, to the special models for a single response variable with several explanatory variables, using the response with the necessary response to explanatory interaction terms. Loglinear models contain same structure as logit models for associations between the explanatory variables and the response variable. To illustrate, for an $I \times J \times 2$ table, it's easy to construct a loglinear model corresponding to the logit model:

$$\log\left(\frac{m_{ij1}}{m_{j2}}\right)=\alpha+\beta_i^C+\beta_j^D. \qquad (1)$$

In (1) the response Y is associated with factors C *and* D, but the effect of each factor is the same at each level of the other factor. The loglinear model has the association terms λ_{ik}^{CY} *and* λ_{jk}^{DY} and the general term λ_{ij}^{CD} for the relationship between the factors. The resulting model will be (CD, CY, DY) .To show that the loglinear model (CD, CY, DY) implies logit model (1), we note that for the loglinear model,

$$\log\left(\frac{m_{ij1}}{m_{ij2}}\right)=\log(m_{ij1})-\log(m_{1j2})$$

$$=[\mu+\lambda_i^C+\lambda_j^D+\lambda_1^Y+\lambda_{ij}^{CD}+\lambda_{i1}^{CY}+\lambda_{j1}^{DY}]$$

$$-[\mu+\lambda_i^C+\lambda_j^D+\lambda_2^Y+\lambda_{ij}^{CD}+\lambda_{i2}^{CY}+\lambda_{j2}^{DY}]$$

$$=(\lambda_1^Y-\lambda_2^Y)+(\lambda_{i1}^{CY}-\lambda_{i2}^{CY})+(\lambda_{j1}^{DY}-\lambda_{j2}^{DY}).$$

For zero-sum constraints $\sum_k \lambda_k^Y=\sum_k \lambda_{ik}^{CY}=\sum_k \lambda_{jk}^{DY}=0$, we have $\lambda_1^Y=-\lambda_2^Y$, $\lambda_{i1}^{CY}=-\lambda_{i2}^{CY}$, *and* $\lambda_{i2}^{DY}=-\lambda_{j2}^{DY}$, since Y has two levels. As a result the logit simplifies to $\log\left(\frac{m_{ij1}}{m_{ij2}}\right)=2\lambda_1^Y+2\lambda_{i1}^{CY}+2\lambda_{j1}^{DY}$.

This is the precise form of the logit model (1), when we identify $2\lambda_{i1}^{CY}$ as the i^{th} effect of C on the logit of $Y(i.e., \beta_i^C)$, $2\lambda_{j1}^{DY}$ as the j^{th} effect of D on the logit of $Y(i.e., \beta_j^D), and\ 2\lambda_1^Y=\alpha$

The λ_{ij}^{CD} terms for association among explanatory variables cancel in the difference in logarithms defined by the logit. The logit model does not contain information about this
association [7].

In summary, when there is a response variable, relevant loglinear models correspond to logit models for that response. When the response has more than

two categories, relevant loglinear models correspond to generalized logit models [7].

2.4 Multiple regression

Multiple regression is a general linear model of the form $y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \varepsilon$ which links a response variable y to several independent (also called explanatory or predictor) variables $x_{1,} x_2, ..., x_{p.}$ This section discusses estimation of model parameters $\beta = (\beta_0, \beta_1, ..., \beta_p)'$ and how to test various hypotheses about them.

Suppose that we have information on n cases, or subjects from $i = 1, 2...., n.$ We may let y_i be the observed value on the response variable and also let $x_{i1}, x_{i2}, ..., x_{ip}$ be the values on the independent or predictor variables of the i^{th} case. The values of the p predictor variables are treated as fixed constants; however, the responses are subject to variability. The model for the response of case i is written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i \qquad (1)$$

$$= \mu_i + \varepsilon_i$$

where $\mu_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$ is a deterministic component that is affected by the regressor variables and ε_i is a term that captures the effect of all other variables that are not included in the model.

We assume that ε_i is a random variable with mean $E(\varepsilon_i) = 0$ and variance $V(\varepsilon_i) = \sigma^2$, and we suppose that the ε_i are normally distributed. Furthermore, we assume that the errors from different cases $\varepsilon_1, ..., \varepsilon_n,$ are independent random variables. All of the above mentioned assumptions imply

that the responses $y_1,...,y_n$ are independent normal random variables with mean $E(y_i)=\mu_i=\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}$ and variance $V(y_i)=\sigma^2$.

The n equations in (1) can also be rewritten in vector form :

$$\begin{bmatrix} y_1 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0+\beta_1 x_{11}+...\beta_p x_{1p} \\ . \\ . \\ \beta_0+\beta_1 x_{n1}+...\beta_p x_{np} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

In short, $y=X\beta+\varepsilon$.The assumptions on the errors in the above model can also be written in vector form. One can write $\varepsilon \sim N(0,\delta^2 I)$, a multivariate normal distribution with mean vector $E(\varepsilon)=0$ and covariance matrix $V(\varepsilon)=\delta^2 I$. On parallel to that we can write $y \sim N(X\beta,\delta^2 I)$, a multivariate normal distribution with mean vector $E(y)=X\beta$ and covariance matrix $V(y)=\sigma^2 I$.

2.4.1 Estimation of the parameters of the model

It is very important to also consider the estimation of the unknown parameters which are; the $(p+1)$ regression parameters β, and the variance of the errors σ^2. Since $y_i \sim N(\mu_i,\sigma^2)$ with $\mu_i=\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}$ are independent then, the joint probability density function can be written as $p(y_1,...,y_n \mid \beta,\sigma^2)$.

If we treat the density function for a given data y ,as a function of the parameters, this leads to the likelihood function:

$$L(\beta,\sigma^2 \mid y_1,...,y_n)=(1/\sqrt{2\pi\sigma})^n \exp\left[-\sum_{i=1}^{n}(y_i-\mu_i)^2/2\sigma^2\right] \qquad (2)$$

If we maximize the likelihood function L with respect to β is equivalent to minimizing $S(\beta)=\sum_{i=1}^{n}(y_i-\mu_i)^2$ with respect to β .The reason being the fact that the exponent in (2) is the only term containing β .The sum of squares $S(\beta)$ can be written in vector notation, as :

$$S(\beta) = (y-\mu)'(y-\mu) = (y - X\beta)'(y - X\beta), \text{ since } \mu = X\beta \quad (3)$$

The minimization of $S(\beta)$ with respect to β is known as least squares estimation, and for normal errors it is equivalent to maximum likelihood estimation. We can determine the least square estimates by obtaining the first derivatives of $S(\beta)$ with respect to the parameters $\beta_0, \beta_1, ..., \beta_p$, and by setting the $(p+1)$ derivatives equal to zero.

The general purpose of multiple regressions is to learn more about the relationship between several explanatory or predictor variables and a response or criterion variable. A multiple regression allows the simultaneous testing and modeling of multiple explanatory variables.

In this study we can also apply multiple regression by taking a behavioral risk factor as a dependent variable, with gender and age as predictor or explanatory variables, together with any other predictor variables.

2.4.2 Multicolinearity

In multiple regressions analysis explanatory variables' multicollinearity is not recommended. In reality, it is very rare to experience multicollinearity in a data set. It can be defined as the situation which arises when some or all of the explanatory variables are so highly correlated one with another that it becomes very difficult, if not impossible, to disentangle their influences and obtain a reasonably precise estimate of their separate effects [25].

The absence of multi-collinearity is essential to a multiple regression model. Collinearity simply means co-dependence**.** According to (Vaughan & Berry, 2005); Collinearity is problematic when one's purpose is explanation rather than mere prediction. Collinearity makes it more difficult to achieve significance of the collinear parameters. But if such estimates are statistically significant, they are as reliable as any other variables in a model, and even if they are not significant, the

sum of the coefficient is likely to be reliable. In this case, increasing the sample size is a viable remedy for collinearity when prediction instead of explanation is the goal (Leahy, 2001). However, if the goal is explanation, measures other than increasing the sample size are needed.

In a regression model containing two explanatory variables x_1 and x_2 which are highly correlated, it is not necessary to include the two of them in the model. This is because the two variables will be expressing the same information, so there is no point to include both. For example, if for any two X variables, X_i and X_j ,we have $r_{ij} = 1.0 \; or \; -1.0$ then the inverse of the symmetric correlation matrix does not exist. The scalar $1/(1-r_{12}^2)$ used in obtaining the inverse of a 2×2 matrix becomes $1/0$, an operation that is not permitted if r_{12} is equal to either $1.0 \; or \; -1.0$.If r_{12}^2 approaches 1.0, and then the scalar $1/(1-r_{12}^2)$ becomes very large.

It is obviously not difficult to recognize a correlation between any two X variables that approaches $-1.0 \; or \; 1.0$, but in large correlation matrices it is possible for one of the X variables, say X_i, to be a linear function of the other remaining X variables such that $R_{i.}^2$,the squared multiple correlation of X_i with the remaining X variables, approaches 1.0.If $R_{i.}^2 = 1.0$, then the inverse of the correlation matrix does not exist.

Evidence regarding a high degree of multicollinearity may be provided by the standard errors of the regression coefficients. When several X variables are involved, we have

$$s_{bi} = \sqrt{\frac{MS_{res}}{\sum x_i^2 (1-R_{i.}^2)}}$$ where $R_{i.}^2$ is the squared multiple correlation of X_i with the remaining X variables. As $R_{i.}^2$ approaches 1, the denominator

becomes small and s_{bi} becomes large. An extremely large value of s_{bi} for any of the regression coefficients may be an indication of multicollinearity

Statistically asset of variables is said to be exactly collinear if there is a linear relationship among the variables. For example we may have $\lambda_1 X_{i1} + \lambda_2 X_{i2} + \ldots + \lambda_k X_{ik} = 0$ for $i = 1, 2, ...n$

where λj are constants and X_j are explanatory variables.

Multicollinearity occurs when the relationship between any one of the explanatory variables and the remaining variables yields a multiple correlation coefficient close to 1.

multicollinearity in a data set can be seen by the presence of the following indicators

- High changes in the estimated regression coefficients when an explanatory variable is added or deleted.
- Statistical significance of an F-statistic for a set of explanatory variables at the same tail as the individual coefficients of the variables fail to exhibit statistical significance

A formal detection–tolerance or variance inflation factor has been suggested by some authors for multicollinearity.

Variation inflation is the consequence of multicollinearity. In a regression model we expect a high variance explained (R-square). The higher the variance explained is, the better the model is. However, if collinearity exists, probably the variance, standard error, parameter estimates are all inflated. In other words, the high variance is not a result of good independent predictors, but a miss-specified model that carries mutually dependent and thus redundant predictors. Variance inflation factor (VIF) is common way for detecting multicollinearity.

Considering the general linear model $y = X\beta + \varepsilon$ with an intercept and p regressors, if we standardize the $y \text{ and } x$ variables,

$$y_i^* = \frac{y_i - \bar{y}}{s_y} \quad and \quad z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad j = 1, 2, ..., p$$

where $\bar{y} \text{ and } \bar{x}_j$ are the corresponding sample means, and $s_y \text{ and } s_j$ are the appropriate sample standard deviations. Hence, the linear model can be expressed as;

$$y^* = \alpha_1 z_1 + \alpha_2 z_2 + ... + \alpha_p z_p + \varepsilon^* \qquad (1)$$

The important point to note is that there are only p regression coefficients and that there is no intercept in (1). The covariance matrix of the least squares estimates of the parameters in the general linear model is $V(\beta) = (X'X)^{-1}\sigma^2$. In the standardized model (1), the matrix that corresponds to $X'X$ reduces to the correlation matrix C. Hence, $V(\hat{\alpha}) = C^{-1}\sigma^2$, where $\alpha = (\alpha_1, ..., \alpha_p)'$. The diagonal elements of C^{-1} are the scaled variances of the least squares estimates, $V(\hat{\alpha}_i)/\sigma^2$. For illustration, we can consider a case when, $p = 2$. Then the model is

$$y^* = \alpha_1 z_1 + \alpha_2 z_2 + \varepsilon^* \qquad (2)$$

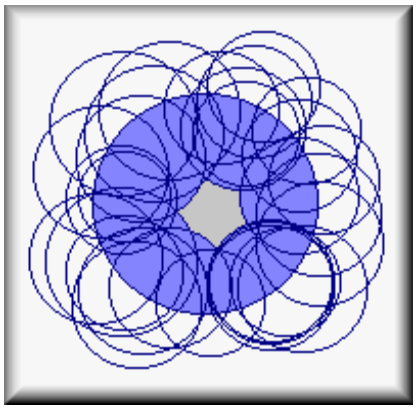
and $$C = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}, with \quad C^{-1} = (1 - r_{12}^2)^{-1} \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}$$

If r_{12} were zero, then C^{-1} has ones in its diagonal, and $\frac{V(\hat{\alpha}_1)}{\sigma^2}, i = 1, 2$ are called variance inflation factors because they measure how the correlation among the regressor variables inflates the variance of the estimates. If these factors are much larger than one there is multicollinearity.

We can illustrate VIF using a Venn diagram in figure 1.
In figure 1, the circle at the center represents the outcome variable and all surrounding ones represent the independent variables. The superimposing area

denotes the variance explained. When there are too many variables, it is likely that Y is almost entirely covered by many inter-related Xs. The variance explained is very high but this model is over-specified and thus useless.

Figure 1: Venn diagram illustrating VIF



tolerance = 1 - R^2, VIF = $\frac{1}{tolerance}$.

R^2 is the multiple correlation coefficient of the explanatory variables as a group, omitting the response variable.

A tolerance of less than 0.20 or 0.10 and/or a VIF of 5 or 10 and higher may indicate a problem of multicollinearity [22].

2.4.3 Consequences of multicollinearity

In the presence of multicollinearity, the estimate of a particular variable's impact on y while controlling for the others tends to be less precise than if there was less correlation between explanatory variables.

The widely used interpretation of a regression coefficient is that it provides an estimate of the effect of a one unit change in an explanatory variable, X_1, holding the other variables constant. If there is a very high correlation between X_1 and another explanatory variable X_2 in a given data set then we only have

observations for which X_1 and X_2 have a particular linear relationship (either negative or positive).

In that circumstance we do not observe situations or cases for which X_1 changes independently of X_2, so we have an imprecise estimate of the effect of independent changes in X_1.

2.4.4 Remedy to multicollinearity

In pure statistical terms multicollinearity does not bias results for fitted values of the response variable, but if there are any other factors whose presence or absence could introduce bias, multicollinearity can multiply the effects of that bias.

Standard use of regression may take coefficients from the model and then apply them to other data. If a relative new data is different from the data that was fitted, this application may introduce large errors in their predictions because the pattern of multicollinearity between the explanatory variables maybe different in new data .

One may drop one of the explanatory variables, to produce a model with statistically significant coefficients. However, the cost might be loss of information because. Leaving out or omitting a relevant variable result in biased coefficient estimates for the remaining explanatory variables [22].

2.5 Previous related research

The main objective of this research is the application of some generalized linear models to show the influence of gender and age on the behavioral risk factors of HIV. In this section we will be examining use of generalized linear models to show the influence of age and gender, in previous related research.

2.5.1 A Study of Gender power imbalance on women's capacity to negotiate self protection against HIV/AIDS in Botswana and South Africa.

This study was carried out during the months of July and December 2003 by Tabitha T. Langen with the help of twelve research assistants, in the Kwazulu Natal province of South Africa and in Botswana, using a sample size of 2658 women aged 18-49 years were surveyed [34].

The two countries were preferred over the countries in Southern Africa because Botswana has the highest prevalence of HIV infection in the world. Some 35.4% of pregnant women in Botswana were found to be HIV positive [23] and South Africa had the largest total count, 5.0 million of adults and children infected with HIV [24].

Logistic regression was one of the statistical methods used to confirm gender and age related influential factors that significantly affect women's ability to protect themselves against the risk of HIV.

The bivariate analysis from Table 4 and Table 5 in *Appendix A* depicted that there were some differences in condom use suggestion by women and condom use refusals by men classified by selected background characteristics.

However, it was not known to what extent the individual variables jointly affected women's ability to suggest condom use to their partners. As a result, it was necessary to use the logistic regression technique to provide a clear perspective on the net effects of the relationship between the dependent and explanatory variables.

One can suggest that a log linear model could be fitted to see whether gender power imbalance with the inclusion of other factors like age was still going to be a significant factor.

When 2 $\times$2 tables apply one may examine the OR's. The study was only carried out in Kwazulu Natal so generalizations cannot be made about the whole of South Africa based only on a study carried out in one province.
Thus, a study which encompasses the whole of South Africa may offer more representative evidence to the effect of gender, age and other factors on behavioral risk of HIV.

The most disturbing revelation of the study was the fact that it was men who had multiple partners who were significantly more likely to refuse to use condoms. This finding can be taken as a challenge to other researchers to find out why these men refuse.
The Kwazulu Natal and Botswana study has shown that there is need to educate women and promote their self confidence.

2.5.2 A cross-sectional study of drug users carried in Pretoria.

A cross-sectional study of drug users was carried out in Pretoria the capital city of South Africa [35]. The data collected was used to model HIV/AIDS infection as a function of sexual risk behaviors and drug use as modified by gender.

Massive gender differences in HIV/AIDS infection exist with females four times as likely to be infected with HIV as compared to their male counterparts [25].

Thus, the main emphasis was to see the effect of gender interactions with age. HIV/AIDS correlations including demographic factors, sexual risk behaviors and drug use were assessed using the logistic regression.

A simple logistic regression was used to calculate unadjusted odds ratios and their 95% confidence intervals, and then multiple logistic regressions were used to compute adjusted odds ratios and their 95% confidence intervals too.

Gender relations and interactions with age were put in the model due to known age gender interactions in the literature. More gender interactions were assessed in terms of risk behaviors using a backward elimination process that included variables for which $\alpha \geq 0.1$

Backward elimination involves starting with all candidate variables and testing them one by one for a discernible statistical contribution, in the process eliminating or deleting any of the variables that do not exhibit statistically significance.

However, the backward elimination process has the following limitations which need to be highlighted:

- A sequential procedure of F-tests is often the main procedure to control the exclusion of variables but these F-tests will be carried on the same data so there will be complications of multiple comparisons for which several correction criteria have been developed.

- It is not easy to interpret the p-values associated with these F- tests since each is conditional on the previous tests of exclusion.

- The tests themselves have an element of bias since they are computed on the same data (Rencher and Pun, 1980, Copas, 1983)

- In contrast Wilkinson and Dalall (1981) computed percentages points of the multiple correlation coefficient by simulation and showed that a final regression obtained by forward selection, reported by the F-procedure to be significant at 0.1% was in fact only significant at 5%.

After the overall assessment of the model several gender interactions were determined to be both statistically significant and clinically significant, therefore a gender stratified analysis was then conducted using two multiple logistic regression models.

The first model to be used was the Overall Model: Multiple Logistic Regression for HIV on demographics as well as drug risk and sexual risk behaviors. Some known interactions incorporating gender and age were incorporated in the model.

A backward elimination process was used to asses gender interactions in terms of sexual risk behaviors, and a final model with a good fit using the Homer and Lemeshow Lack of Fit Test (chi-square(df)=8.59(8), $p-value$ = 0.38) and a global null hypothesis test (chi-square(df)=88.44(23), $p-value$<0.0001) had the interaction terms such as gender and age included (Wald chi-square(df)=9.86(2), $p-value$<0.01).

In contrast a Stratified Model: Multiple Logistic Regression exhibited several gender interactions which were used to compare the adjusted odds of HIV/AIDS for males against females.

Thus, a multiple logistic regression model was estimated separately for males and females. Same covariates as the ones used in the overall model were used also in the two stratified models, giving a global fit (Chi-square (df) =43.69(16), $p-value$<0.001) for males. For females, the global fit statistics test (Chi-square (df) =17.46(16), $p-value$=0.42) indicated that the covariates were not adequate in explaining HIV odds for females.

Adjusted odd ratio results indicated that amongst males older age was associated with testing positive for HIV. However, the same strong association was not demonstrated amongst females. Gender stratified analysis further revealed that males who used condoms in their last sexual activity were less likely to test positive for HIV (Adjusted Odd Ratio=0.37, 95%Confidence Interval=0.14, 0.98, $p-value$ < 0.05). However the same association was not observed in females.

The Pretoria study suggests that gender and age play some roles on the behavioral risks of HIV/AIDS. The statistical methods used in the research included the logistic regression and multiple regressions with some backward elimination techniques.

Some drawbacks of the study include the limited sample size for tests of interaction, which could not enable further delineation of gender differences. Finally, it should be noted that a particular overall model gave a best fit but once the analysis was stratified by gender the same model only produced a best fit for males but not females. Future studies may need to put their focus on the predictors of HIV/AIDS in females

2.5.3 A Study of effectiveness of an HIV prevention intervention for African American women at Virginia Commonwealth University.

Women were recruited from three local colleges and universities and several community-based agencies including health clinics, faith based institutions in a southeast metropolitan area in Virginia USA [36].

The main purpose of the study was to OBTAIN a thorough understanding and to examine unique contribution of age to HIV risk and protective behaviors. Variables that have been found to be associated with HIV risk and protective behaviors were controlled, such as level of education, relationship length and partner status.

There were four hypotheses of the test namely:

- Age was associated with condom use after controlling for education, partner status, and relationship length, in that as age increases condom use and condom use intentions will decrease

- Age was associated with condom attitudes in that as age increases positive attitudes towards condoms will decrease.
- Age was associated with condom negotiation efficacy and as age increases condom use efficacy will decrease
- Age was associated with HIV knowledge in that as age increases HIV knowledge will decrease.

The explanatory variable of the study was age whilst there were 3 control variables, namely education, partner status and length of relationship. Dependent variables included HIV risk and protective behaviors. A number of variables were used to assess HIV risk and protective behaviors

Hierarchical multiple regression and logistic regression were the statistical methods used to test the study hypotheses. The analysis mainly investigated the effect of age on HIV risk and protective variables after controlling for the effects of education, partner status, and relationship length.

After controlling for the effects of education, partner status and relationship length, age was found to be significantly associated with the percentage of persons using condoms at last engagement within previous 3 months.. The F – test F (1,180) = 4.47, $p-value<0.05$, condom use at last sexual engagement, $\chi^2(4)$=32.98, $p-value<0.01$ and condom use intentions, $\chi^2(4)=40.63, p<0.01.$ It was noted that as age increased, both condom use and the intentions to use condoms decreased.

The second hypothesis was that as age increased, positive attitudes toward condoms would decrease. However, age was not significantly associated with condom attitudes (1, 289) = 0.11, $p-value$ > 0.05. One of the control variables, level of education, had a small effect on condom attitudes.

The third hypothesis predicted that as age increased, condom negotiation efficacy and condom use efficacy would decrease. It was found that age was significantly associated with condom negotiation efficacy, F (1,272) =5.07, $p-value$ < 0.05, but in an opposite direction towards what was predicted, as the age increased condom negotiation efficacy increased too.

Age was found to have a little effect on condom negotiation efficacy and it accounted for a paltry statistically insignificant 2% of the variability in condom negotiation efficacy after controlling for the effects of education, partner status and relation ship length. Age did not predict condom use efficacy, F (1,295) =2.59, $p-value$ >0.05.

The fourth hypothesis predicted that as age increased, HIV knowledge would also decrease. It was found that age was not significantly associated with HIV knowledge F (1,281) = 0.26, $p-value$ > 0.05.

From the study, it was inferred that after controlling other factors like education, length of relationship and partner status, age showed small effects on condom use. Similarly, younger women as compared to older women were more likely to report to using condoms in their current and past sexual relationships.

As age increased participants reported that their partners had less favorable attitudes toward condom use; this suggests that it may be important to include partners in interventions to address building more positive attitudes towards using condoms in relationships.

In contrast to study hypotheses, there was evidence that as age increased women reported higher efficacy for negotiating condom use with a partner. This study supports the targeting of current prevention efforts that use skills building approach to teach young women assertiveness and condom negotiation skills.

The study has some limitations including the fact that age categories involved unequal numbers of women, and consequently it was difficult to conduct age analyses by cohorts.

From the study we may note that as the age of participants increased, participants reported less frequent condom use, reported lower condom use intentions, and perceived their partners attitudes toward condoms to be less favorable, even after controlling other variables which might have an influence. As, age increased, participants showed a higher condom negotiation efficacy.

The statistical methods used which are multiple regression and logistic regression.

2.5.4 Predictors of Condom Use among Young Adults in South Africa: The Reproductive Health and HIV Research Unit National Youth Survey

This study was carried out in 2003 by the Reproductive Health and HIV Research Unit (RHRU) at the University of the Witwatersrand in South Africa. A sample of 7686 sexually experienced young adults aged 15-24 were randomly recruited [37].

The statistical method applied in this study, was a multiple logistic regression model examining predictors of condom use by gender. The results of the model showed that male respondents who had used a condom at their sexual debut were almost 6 times as likely to have used condom during their most recent sexual intercourse as those who had not used a condom at their sexual debut. (odds ratio [OR]=5.92 ;95% Confidence interval[CI]=4.02,8.72)

However, the study had some limitations such as over representation of black Africans, even though the study was intended to be nationally representative.

2.6 Conclusion

In several previous sources and studies, it is clear that gender and age had an influence or impact on the behavioral risk factors of HIV. Gender plays a pivotal role in shaping the opportunities one is offered in life, the roles one may play, and the kinds of relationships one may engage in.

From the literature review it has been shown that the logistic regression is a dominant statistical method used to show the influence of age and gender on the behavioral risk factors of HIV.

Even though the behavioral risk factors of HIV affect both men and women, it seems women are the ones who are more vulnerable to these factors because of biological, social, cultural and economic factors. Thus, existing efforts being implemented by the government and non-governmental organizations may need to be reviewed from a gender sensitive perspective to ensure positive and sustainable changes in behavioral risk.

CHAPTER THREE

Research methodology

3.1 Introduction

This chapter describes the practical elements and the conceptual framework of the research. It explores the research question in more depth, and describes which methods are more suitable, given the nature and objectives of the research.

The main generalized linear models which will be applied on this research include logistic regression, multiple regression and log linear models. The main statistical software that will be used in this study is the SAS software system.

3.2 Logistic model

As already stated in chapter two, a logistic model is a type of explanatory or predictive model for a dichotomous response variable .

3.2.1 Fitting the Logistic model

We can use the LOGISTIC procedure, GENMOD procedure and the CATMOD procedure to fit a logistic regression model. PROC LOGISTIC is specifically designed for logistic regression. The LOGISTIC procedure has the capabilities for stepwise, forward, backward, and/or selection of best subset of explanatory variables among multiple explanatory variables.

However, it is important to note that PROC LOGISTIC requires all data cases to be complete and does not predict the estimated linear predictor and its standard error estimate, the fitted probabilities, and confidence limits, and the regression diagnostic statistics for any observation with missing explanatory variable values.

By default any observation with missing values for the response or explanatory variables is excluded from the analysis. Therefore, we may first estimate data for missing values by using PROC MI in SAS.

PROC GENMOD is a procedure which was introduced in SAS for fitting generalized linear models. It uses a class statement for classifying categorical variables, so indicator or dummy variables do not have to be constructed in advance.

The CATMOD procedure is a general procedure designed to fit models to functions of the frequencies obtained for categorical response variables.

In our study we will mainly use the PROC LOGISTIC procedure for the fitting of the logistic regression. We will be using the three main approaches which are:

- ✓ Forward selection – which involves starting with no explanatory variables in the model, trying out the variables one by one and including the variable with most discernible effect provided that it is statistically significant.
- ✓ Backward elimination – which involves starting with all candidate variables and testing them one by one for statistical significance, and deleting the variable with least discernible effect, provided that it is not statistically significant .
- ✓ Stepwise selection - which combines the elements of the previous two.

3.2.2 Goodness of fit of the Logistic model

The goodness of fit of a statistical model describes how well the model fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question.

The purpose of an overall goodness of fit test is to determine whether the fitted model adequately describes the observed outcome experience in the data. One

concludes a goodness of fit if the differences between the observed and fitted values are small and if there is no systematic contribution of the differences to the error structure of the model.

3.2.3 The Hosmer and Lemeshow chi-square(H-L chi-square) test of goodness of fit.

The H-L chi-square is the recommended test for the overall fit of a binary logistic regression model though it is not available in multinomial logistic regression. The H-L chi-square divides subjects into deciles based on predicted probabilities and then computes a chi-square from observed and expected frequencies.

Then, a probability value is computed from the chi-square distribution to test the fit of the logistic model. If the p-value of the Hosmer and Lemeshow goodness of fit test statistic is greater than 0.05, as we want for well fitting models, we do not reject the null hypothesis that there is no difference between observed and model predicted values. The model estimates fit the data at an acceptable level.

3.2.4 Pearson and deviance goodness of fit

The Pearson statistic is based on traditional chi-square for categorical variables and the deviance statistic is based on likelihood ratio chi-square. The deviance test is preferred over the Pearson [27].

3.2.5 The likelihood ratio test

The likelihood ratio test is also called the log-likelihood (LL) test, and is based on the variance -2LL. The likelihood ratio test is a test of the significance of the difference between the log likelihood for the research model and the log likelihood ratio for a reduced model. This difference is called the reduced model chi-square. The likelihood ratio test is generally preferred over its alternative, the Wald test, discussed below.

3.2.6 The Wald statistic (test)

The Wald statistic is a commonly used test of significance for individual logistic regression coefficients of each explanatory variable. The researcher may want to drop an explanatory variable from the model when its effect is not significant by the Wald statistic.

According to [27] for large logit coefficients, the corresponding of standard error is inflated, lowering the Wald statistic and leading to type ll errors. As a result there is a flaw in the Wald statistic in that very large effects may lead to large standard errors and small Wald chi-square values. In consequence a strong explanatory variable may be omitted. The like hood ratio approach is preferable because of that reason.

3.3 Information theory measures of model fit

- The Akaike's information criterion (AIC) is a common information theory statistic used when comparing alternative models. It is calculated as -2log likelihood + 2k where k is the number of estimated parameters. It is, reported in PROC LOGISTIC output.
- The Bayesian Information criterion (BIC) is a common information statistic also used when comparing alternative models. A Lower BIC gives a better model.

The Schwartz criterion (SIC) is a modified version of AIC and is part of the PROC LOGISTIC output. Compared to AIC, SIC penalizes over-parameterization more which in turn rewards model parsimony. A Lower SIC gives a better model. The SIC can be calculated as -2log likelihood + k log n, where n is the sample size of the study.

3.4 Loglinear model

The loglinear models can be used to analyze the relationship between two or more categorical variables (The data are summarized as two-way or multiway contingency tables of frequencies).

3.4.1 Fitting the Loglinear model

Once a model has been chosen for a frequency table of categorical variables ,the expected frequencies need to be obtained. Fitting a log linear model is a process of deciding which associations are significantly different from zero; the corresponding interaction terms are included in the final model used to explain the observed frequencies.

Terms which are excluded from the model go into the residual or error term, which reflects the overall lack-of-fit of the model. The usual goal of log-linear modeling is to find a small model (few terms) which nonetheless achieves a reasonable fit (small residuals or a small lack of fit statistic).

Loglinear models can be fitted using PROC CATMOD and PROC GENMOD. The iterative proportional fitting process generates maximum likelihood estimates of the expected cell frequencies for a hierarchical model.
Once estimates of the expected frequencies for the given model are obtained, these numbers are entered into appropriate formulas to produce the effect parameter estimates for the variables and their interactions in the current model [28].

For larger tables an iterative proportional fitting algorithm (Deming-Stephan algorithm) is used to generate expected frequencies. This procedure uses marginal tables fitted by the model to insure that the expected frequencies sum across the other variables to equal the corresponding observed marginal tables.

3.4.2 Goodness of fit of the loglinear model

After fitting the model, it is very important to make a decision about which particular model provides the best fit. The overall-goodness of fit of a model can be assessed by making comparisons on the expected frequencies to the observed cell frequencies for each model.

Basically, we use the Pearson chi-square test statistic or the Likelihood ratio (L^2) to test a models fit. In most cases L^2 is used in maximum likelihood estimation. The formula for the L^2 statistic is as follows:

$$L^2 = 2\sum n \ln\left(\frac{n}{m}\right)$$

where n and m denote the observed and fitted frequencies respectively and summation is over all cells in the contingency tables. Larger L^2 values indicate that the model does not fit the data well and thus the model should be rejected [29].

It is usually found that more than one model provides an adequate fit to the data as indicated by the non-significance of the likelihood ratio. The likelihood ratio to compares an overall model nested within a smaller model (i.e. comparing a saturated model with one interaction or main effect dropped to assess the importance of that term).

Some limitations of the loglinear model involve the inclusion of many variables in loglinear models often making interpretation very difficult. Also, the metric requires a large sample before an adequate fit can be attained by a loglinear model.

3.5. Multiple regression

A multiple regression allows the simultaneous testing and modeling of multiple explanatory variables.

3.5.1 Fitting a multiple regression model

Before we fit a multiple regression model in SAS we need to examine the correlations among the predictor variables and dependent variables using PROC CORR. We can first use the default settings from PROC CORR, which gives us a correlation matrix with pair wise deletion of missing values.
This will even enable us to note correlations amongst variables and occasionally even some multicollinearity [30].

Next we examine the correlations using the NOMISS option, which gives us a correlation matrix after list wise deletion of cases. That is, only those cases that have complete data for all variables will be included in the correlation matrix. We can even plot a scatter plot for the predictors and finally fit the multiple regression.

In this study we have with the chosen behavioral risk as a dependent variable and then take age and gender as the predictor variables.
Various options of regression which can be used in fitting a multiple regression model have been cited as:

- Stepwise regression
- Forward selection
- Backward elimination

3.5.2 Goodness of fit of a multiple regression model

The fit of the multiple regression model can be assessed by the Coefficient of Multiple determination, which is a fraction that represents the proportion of total variation of the dependent variable that is explained by the regression plane.

Sum of squares due to error (SSE) = $\sum_{i=1}^{n}(y_i - \hat{y}_i)$.

Sum of squares due to regression (SSR) = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$.

Total Sum of squares (SST) = $\sum_{i=1}^{n}(y_i - \bar{y})^2$.

The ratio SSR / SST denotes the proportion of the total variation in the dependent variable explained by the regression model.

This ratio represented by R^2, is called the coefficient of multiple determination. R^2 is sensitive to the magnitudes of n and k, where n and k represent number of observations and number of predictors respectively. If k is large relative to n, the model tends to fit the data very well, with large R^2 values. In extreme cases, if $n = k+1$, the model would exactly fit the data, and R^2 =1.

However there is a better goodness of fit measure namely adjusted R^2, which can be calculated as Adjusted R^2 = $1-\left(\frac{n-1}{n-k-1}\right)(1-R^2)$

= $1-\frac{SSE/(n-k-1)}{SST/(n-1)}$.

The overall goodness of fit of the regression model can be evaluated using an F-test in the format of analysis of variance.

Under the null hypothesis: $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$, the F-statistic $\frac{[SSR / k]}{[SSE / (n-k-1)]} = \frac{MSR}{MSE}$ has an approximate F -distribution with k and $n-1$ degrees of freedom.

For any specific variable x_i we can test using the null hypothesis $H_0 : \beta_i = 0$, by computing the statistic $t = \frac{b_i - 0}{SE(b_i)}$ and performing a one or two tailed $t\ test$ with $n-k-1$ degrees of freedom.

3.6 Conclusion

This chapter has presented the main different statistical methods and tools that can be used to exhibit the influence of gender and age on the behavioral risk factors of HIV. These statistical methods will be implemented in this study to achieve the main objectives of the research.

CHAPTER FOUR

Data analysis and results

4.1 Introduction

This chapter describes and explains methods and applications through a thorough detailed discussion of the outputs of the results and their interpretations.

In this chapter we will fit the three statistical methods, namely logistic regression, loglinear regression and multiple regression. After fitting each and every model, there will be the analysis of the results for that model. The main objective is to find a parsimonious version for each statistical model.

Section **4.4** will explain the analysis of the logistic regression and its outputs. Again, section **4.5** will discuss the analysis of the loglinear regression. Multiple regression will be discussed in section **4.6.** Section **4.7** is the conclusion.

4.2 Methods of data collection

This study uses primary data, we collected the data ourselves. There are many different methods of collecting primary data and the main methods include questionnaires, interviews, case studies, diaries and observations. We used a questionnaire. A questionnaire can be defined as a series of questions asked to individuals to obtain statistically useful information about a topic of interest. Using the questionnaire enables us to create variables by selecting specific questions from the questionnaire. Every question in the questionnaire is part of the eventual analysis.

This study was carried out in the University of Fort Hare situated in Alice, Eastern Cape, South Africa in July 2010. Fort Hare is a black university with a Student Population of approximately 10 000. The majority of fort hare students are South Africans with approximately 15% of the student population being students from neighboring African countries.

The study was carried among first year students at the University Of Fort Hare, whereby a sample size of 210 students were surveyed though the needed sample size was 200 and an extra of ten students was added in case some students withdrew from the study. I the researcher conducted the research with the assistance of my fellow classmates and supervisor. The main reason the target population were first years is because they were likely to give the correct information without any bias and it was also easy to work with them.

A sample size of 200 was chosen after cost considerations (e.g., maximum budget, desire to minimize cost), administrative concerns (e.g., complexity of the design, research deadlines) and the required acceptable level of precision were attentively scrutinized. The sampling technique used in this study was simple random sampling. The university registry provided the whole list of first year students enrolled in 2010 and simple randomization was used to get the 210 students. Posters, University emails and flyers were used for informing the selected students and where to meet the researcher for more information about the study.

There are many different methods of collecting primary data and the main methods include questionnaires, interviews, case studies, diaries and observations. From the above mentioned methods of data collection, we used a questionnaire. A questionnaire can be defined as a series of questions asked to individuals to obtain statistically useful information about a given topic. Using the questionnaire enables us to create variables by selecting specific questions from the questionnaire. In this study, pre-testing the questionnaire was first done to

ensure that the questions were not confusing or ambiguous, potentially offensive to the respondent leading to biased responses. After pre-testing a final questionnaire was printed out and it can be found in the appendix.

As the owner of the study I thought about the risk behaviors which affect University students and even put them at high risk of HIV/AIDS. There were a lot of behaviors which came to my mind including having unprotected sex, having many sexual partners etc. This helped me to easily design my questionnaire as in my mind already had some questions to ask in order ot address and find the main factors affecting students' behavior

It is important to note that the study was conducted after a clearance letter was obtained from the University of Fort hare's ethical committee. There were no risks involved for students in being part of the study other than some personal questions. The main ethical challenge was the fact that there were questions in the survey which included personal issues like sexual behavior. In pre-testing it was noticed that some students left the section unanswered maybe due to some embarrassments or sensitivity of the questions.

This challenge was addressed by letting students know that the information provided in the study will be kept private and strictly confidential. Also, students were promised to be offered counseling in some cases if needed. It is also important to mention that we were dealing with new students at the university so some issues were still embarrassing for them.
A clearance letter from the University's ethical committee can also be found in the appendix.

After, collecting the data the next step was data analysis AS, SPSS and Stata were the three statistical softwares to be used for data analysis. Before conducting data analysis the data was first checked of its accuracy through data screening. In data screening quantitative variables were examined for the range

of values to be sure that no cases had values outside the range of possible values. The second case was to deal with missing data in the study. There were few cases of missing data in the study. A dichotomous dummy variable was created in the study coded, so that one group includes cases with values on a given variable and the other group contains cases with missing values on that variable. A simple independent samples t- test was run to examine, if there were statistically significant differences in responses between the two groups. In this study there were no significant differences and even the cases with missing data were very few so they were deleted from the data. Also, the data was very complete and the important information needed was provided on the questionnaire to the satisfaction of the researcher.

The main emphasis of the study the statistical methods that can be used to exhibit the influence of age and gender on behavioral risk factors of HIV. Several generalized linear models which will be used. We first describe the variables that will be used in this study.

4.2.1 Variables

The main objective of the study is to assess the statistical methods, which can be used to model the influence of age and gender on the behavioral risk factors of HIV/AIDS. In this study we decided to have six variables though the main predictor variables are gender and age. The six variables are behavioral risk(BHR),age(AGE),gender(GEN),alcohol and drug abuse(ALCD),number of partners(NUMP) and religious beliefs(RELB). The variables were extracted from specific questions in the questionnaire.

The response variable BHR was constructed by finding the mean of 24 responses from the questionnaire which were related to risk behavior of HIV/AIDS. This can be shown as follows:

BHR=(SXATT1+SXNRM1+PSXRNM1+SXINT1+RSXFPEE1+RSXMPEE1+SXBHBLF1+RNSNOUT1+RNSBRKU1+ABBHBLF1+HEDON1+PREVEN1+PREACT1+BEHBLF1+IMPULS1+TSKBLF1+RCDBLF+NEGBLF1+AVIL1+EFFSCOR1+NCBHBLF1+CDMATT1+SXNRM1+CDMINT1)/**24**

Each and every respondent had a BHR value which was considered as risky behavior if the mean score was at least 2 and no risky behavior otherwise. The response variable in this study was considered as a dichotomous dependent variable which was coded as 1 for the presence of BHR and 0 for the absence of BHR. This was done in order to transform the response variable into a categorical variable to enable logistic and loglinear regression to be easily conducted.

The other five predictor variables are ordinal, categorical and continuous. The first predictor variable is AGE which can be treated as an ordinal or continuous variable. AGE was part of the questions which were asked the respondents so it was easy to get the variable from the responses. The second predictor variable is GEN which can be treated as a categorical variable. GEN was also easily accessible as the respondents were asked their gender in the questionnaire.

The last three explanatory variables which are ALCD, NUMP and RELB are ordinal variables. The three predictor variables were also obtained from the research questions in terms of the mean scores of specified questions from the questionnaire as follows:

ALCD=(CAGE11+CAGE21+CAGE31+CAGE41)/**4**;
NUMP=(SX3MO1+SX3MCD1+OSX3MO1+OPTSX1+OSX3MCD1)/**5**
RELB=(ZCHOFTN1+ZBIBLE1+ZMUSIC1+ZRELRAD1+ZRELTV+ZGRACE1+ZPRAY1)/**7**

However, it is important to note that a variable may be transformed from being continuous to categorical or ordinal depending on the statistical method being used. Since, we have already highlighted that the main explanatory variables in the study are AGE and GEN; three more explanatory variables which were

thought to be important in the study were added to even examine the behavior of the response variable and even some interactional effects. The six variables to be used in the study can be clearly listed as follows:

- BHR (Behavioral risk)
- AGE (Age)
- GEN (Gender)
- ALCD (Alcohol and drug abuse)
- NUMP (Number of partners)
- RELB (Religious beliefs)

The variables used in the study were assumed to be having the same weight though in reality, it is very difficult to find a situation where all the explanatory variables have the same impact on the response variable. The main reason in the study the explanatory variables were assumed to be of the same weight was that all of them were considered very important in determining the risky behaviors of students.

SAS Code 4.1

```
BHR=(SXATT1+SXNRM1+PSXRNM1+SXINT1+RSXFPEE1+RSXMPEE1+SXBHBLF1+RNSNOUT1+R
NSBRKU1+ABBHBLF1+HEDON1+PREVEN1+PREACT1+BEHBLF1+IMPULS1+TSKBLF1+RCDBELF
+NEGBLF1+AVIL1+EFFSCOR1+NCBHBLF1+CDMATT1+SXNRM1+CDMINT1)/24
AGE=AGE;
GEN=GENDER;
ALCD=(CAGE11+CAGE21+CAGE31+CAGE41)/4;
NUMP=(SX3MO1+SX3MCD1+OSX3MO1+OPTSX1+OSX3MCD1)/5
RELB=(ZCHOFTN1+ZBIBLE1+ZMUSIC1+ZRELRAD1+ZRELTV+ZGRACE1+ZPRAY1)/7
```

4.3 Logistic regression Analysis

In logistic regression the response variable behavioral risk will be modeled against the predictor variables. Backward elimination, forward selection and stepwise selection are the approaches applied. In this model we will use backward elimination in which all predictor variables are first included in the

model, and then only those that are significant will be retained in the model until a parsimonious model is obtained.

In simpler terms, this means that a saturated model will be fitted first. After that, only those predictors which are significant will be fitted. This will be done until there are no changes in the predictor variables and a best model will be chosen. The SAS code used to model the logistic regression is shown below:

SAS Code 4.2

```
proc logistic;
model BHR=AGE GEN ALCD NUMP RELB;
run;
```

The output of the above SAS code is shown below and it consists of four parts namely:

- Model fit Statistics
- Testing Global Null Hypothesis
- Analysis of Maximum Likelihood Estimates
- Odds ratio Estimates

Output Logistic Regression

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	237.909	226.443
SC	241.212	246.263
-2 Log L	235.909	214.443

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	21.4659	5	0.0007
Score	20.6203	5	0.0010
Wald	18.2854	5	0.0026

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	4.4864	2.7574	2.6472	0.1037
AGE	1	-0.0341	0.1137	0.0901	0.7641
GEN	1	-1.3342	0.4188	10.1491	0.0014
ALCD	1	-0.1566	0.5468	0.0820	0.7746
NUMP	1	-0.7198	0.1942	13.7343	0.0002
RELB	1	0.0621	0.2603	0.0570	0.8113

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
AGE	0.966	0.773	1.208
GEN	0.263	0.116	0.598
ALCD	0.855	0.293	2.497
NUMP	0.487	0.333	0.712
RELB	1.064	0.639	1.772

4.4.1 Model fit statistics

From the results we can see that the model fit statistics, which lists the information theory measures of model fit, which are Akaike Information Criterion(AIC), Schwarz Criterion (SIC) and negative of twice the log likelihood(-2logL). The AIC and SIC can be used to compare different models and the ones with low values are preferred.

In our model the AIC, SC and -2IogL are having lower values and this shows a good fit for the model.

4.4.2 Testing global null hypothesis

The null hypothesis tested is Beta=0, where Beta comprises the parameter vector of regression coefficients

There are three chi-square statistics, namely Likelihood ratio, Score and Wald with the values 21.4659; 20.6023; and 18.2854 respectively, testing the hypothesis that all the predictor variables have coefficients of zero. Since the statistics are significant as their p-values are less than 0.05, we therefore reject the null hypothesis and conclude that at least one of the coefficients is not zero.

4.4.3 Analysis of maximum likelihoods

The "Analysis of Maximum Likelihood Estimates" section lists the parameter estimates, their standard errors, and the results of the Wald test for individual parameters

Here we see the significant variables in the model are GEN and NUMP, as evidenced by P- values which are less than 0.05.
The binary logistic model can be represented as follows:

$$\ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = 4.4864 - 1.3342GEN - 0.7198NUMP\,.$$

4.4.4 Odds Ratio Estimates

The last section of the results gives the odds ratio estimates which measure the strength and direction of association of the variables.

The variable GEN measures the effect of gender on the behavioral risk of HIV, and it shows that males are 0.263 times likely to be involved in behavioral risks as compared to their female counterparts. The same criteria can be used to explain the odds ratios of the other variables.

After backward elimination the variables GEN and NUMP will be kept in the model and the rest will be discarded as they are not statistically significant.

The SAS code to fit the two variables yields the following output is:

SAS Code 4.3

```
proc logistic;
model BHR=GEN NUMP;
run;
```

Output 4.2

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	237.909	220.655
SC	241.212	230.565
-2 Log L	235.909	214.655

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	21.2536	2	<.0001
Score	20.4088	2	<.0001
Wald	18.1066	2	0.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	3.6256	0.7378	24.1449	<.0001
GEN	1	-1.2648	0.3825	10.9341	0.0009
NUMP	1	-0.7328	0.1901	14.8577	0.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
GEN	0.282	0.133	0.597
NUMP	0.481	0.331	0.698

The output was obtained by backward elimination of the logistic regression and the two variables GEN and NUMP were both significant again. Our main explanatory variables of interest were gender and age and from the logistic regression outputs, it was found that gender is the only one which had a significant impact on behavioral risk of HIV/AIDS. The age categories for age were above 20 years and at most 20 years of age. However, in the questionnaire age was given as a number but the researcher categorized the data.

We also tried to fit the model using the two variables of main interest and still it was gender only which was found to be significant. The main reason we chose

the two explanatory variables of main interest is because we had exhausted all our variables using the backward elimination regression procedure.

It is of paramount importance to note that this study was carried among university students so many of them are in the same age category so that might have played a role on its own. In simpler terms, all the respondents of the study were first years and most of them were of the same age groups so their behaviors were likely to be similar also.

The model below can be described as the model of best fit with the two variables both significant:

$$\ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = 3.6256 - 1.2648GEN - 0.7328NUMP\ .$$

4.5 Analysis using Loglinear

Next we will go to the next statistical method which is loglinear analysis and try to fit the same data using a log linear model. The main aim or goal of loglinear modeling is to find a small model which nonetheless achieves a reasonable fit.
As already mentioned above that our variables consists of BHR as our response variable and the predictor variables AGE, GEN, ALCD and NUMP, RELB
However, the variables investigated by loglinear models are all treated as response variables. In other words, no distinction is made between response and predictor variables. Firstly, a SAS code will be used to fit all predictor variables and from the output we will discard all those interactional effects which are not significant until a parsimonious model is obtained.

Previously fitting the l model we noted that the variables AGE, ALCD and RELB were not statistically significant However, since our main variables of interest are age and gender, we cannot discard AGE in the loglinear model.

We will fit all the variables mentioned except RELB and ALCD. The procedure to be used will be the PROC CATMOD in SAS. In this case we will start with a saturated model and begin to delete non significant higher order interaction terms until the fit of the model becomes unacceptable. The SAS code below shows the catmode procedure used to fit the loglinear model.

SAS Code 4.4

```
proc catmod data=tlou;
weight count;
model BHR*AGE*GEN*NUMP=_response_;
loglin BHR|AGE|GEN|NUMP;
run;
```

The corresponding output follows.

Output 4.3

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
BHR	1	26.71	<.0001
AGE	1	18.98	0.7870
BHR*AGE	1	0.01	0.9249
GEN	1	0.10	<.0001
BHR*GEN	1	4.09	0.0431
AGE*GEN	1	13.28	0.0003
BHR*AGE*GEN	1	0.04	0.8377
NUMP	2	7.34	0.0255
BHR*NUMP	1*	0.75	0.3854
AGE*NUMP	1*	0.33	0.5683
BHR*AGE*NUMP	1*	0.07	0.7972
GEN*NUMP	1*	4.44	0.0352
BHR*GEN*NUMP	1*	0.10	0.7482
AGE*GEN*NUMP	1*	0.24	0.6237
BHR*AGE*GEN*NUMP	1*	0.04	0.8356
Likelihood Ratio	0	.	.

4.5.1 Analysis of Maximum Likelihood Analysis of Variance (Output 4.3)

The likelihood ratio test of fit is zero for the saturated model with zero degrees of freedom.

We also observed that four factor and three way interaction terms have large p-values leading to their elimination because of non significance.

Similar SAS code consisting of three way interactions estimating a model with the exclusion of the four way interaction terms with none of the three way interaction terms being significant.

As a result we considered a unique model which totally eliminates and excludes all the three-way and four way interactions as shown below:

SAS Code 4.5

```
proc catmod data=tlou;
weight count;
model BHR*AGE*GEN*NUMP=_response_;
loglin BHR|AGE BHR|GEN BHR|NUMP AGE|GEN AGE|NUMP GEN|NUMP;
run;
```

The corresponding output follows:

Output 4.4

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
BHR	1	31.59	<.0001
AGE	1	21.22	0.6369
BHR*AGE	1	0.05	0.8204
GEN	1	0.22	<.0001
BHR*GEN	1	4.88	0.0271
NUMP	2	10.87	0.0044
BHR*NUMP	1*	0.56	0.4542
AGE*GEN	1	20.35	<.0001
AGE*NUMP	1*	0.88	0.3472
GEN*NUMP	1*	7.00	0.008
Likelihood Ratio	5	0.67	0.9844

4.5.2 Analysis of Maximum Likelihood Analysis of Variance (output 4.4)

The above output 4.4 depicts a better model as compared to the saturated model as we can see that likelihood ratio is 0.67 and 5 degrees of freedom and a p-

value of 0.9844. We have six of the Wald tests being statistically significant, with the p-value being less than 0.05.

Another model was fitted by excluding AGE, BHR*AGE and BHR*NUMP. The reason we excluded the two way interactions is because of their high p-values.

SAS code 4.6

```
proc catmod data=tlou;
model BHR*AGE*GEN*NUMP=_response_;
loglin BHR|GEN AGE|GEN GEN|NUMP;
run;
.
```

The corresponding SAS code follows:

Output 4.5

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
BHR	1	38.39	<.0001
GEN	1	2.40	<.0001
BHR*GEN	1	5.01	0.0252
AGE	1	25.86	0.1211
AGE*GEN	1	19.78	<.0001
NUMP	2	33.97	<.0001
GEN*NUMP	1	3.23	0.0725
Likelihood Ratio	5	9.45	0.0923

4.5.3 Analysis of Maximum Likelihood Analysis of Variance (output 4.5)

The above output 4.5 depicts that this is the best model that can be obtained from the fitted data. So after dropping the two way interactions BHR*AGE and BHR*NUMP we noticed that the likelihood ratio rose from 0.67 to 9.45 and the p-value also dropped to 0.0923.

All the two way factor interactions were significant with the exception of GEN*NUMP though its also very close to being significant since its p-value is 0.0725. Loglinear models are models that postulate a linear relationship between the independent variables and the logarithm of the dependent variable for example $\log(y) = a_0 + a_1x_1 + a_2x_2 .....a_Nx_N$

where y is the response variable; x_i, i =1… K are explanatory variables, and {a_i, i =0… N} are parameters (coefficients) of the model.

Loglinear models, for example, are widely used to analyze categorical data represented as a contingency table.

In this case, the main reason to transform frequencies (counts) or probabilities to their log-values is that provided the explanatory variables are not correlated with each other, the relationship between the new transformed response variable and the explanatory variables is a linear (additive) one. From output 4.5 the interactional effects between age and the response variable were statistically significant and this further justifies the significance of the explanatory variable gender. For example, a simple bivariate independence model for two categorical variables $X \ and \ Y$

$p_{ij} = P(X = i)\ P(Y = j); i = 1,....,M;\ j = 1,.....,N$. This transforms to

$\log(p_{ij}) = \lambda_i^X + \lambda_j^Y$; where $\lambda_i^X = \log P(X = i)$; $\lambda_j^Y = \log P(Y = j)$.

In our case we have four variables so the model can be transformed as below.

One can conclude that the parsimonious loglinear model for our data is:

$$\ln(m_{ijkl}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_{ij}^{AB} + \lambda_{jk}^{BC} + \lambda_{jl}^{BD},$$

where A, B, C, D denotes BHR, GEN, AGE and NUMP respectively as explained in the hierarchical modeling of the loglinear model in chapter 3

If we examine model we can see that there is a significant association between gender and behavioral risk. Also there is a significant association between gender and age. Since, in our study we wanted to show the influence of age and gender on the behavioral risk, one can say the loglinear model has also provided evidence to show that only one of the two variables have got an effect on the behavioral risk factors of HIV/AIDS.

4.6 Analysis using Multiple Regression

A third method used to show influence of age and gender on the behavioral risk of HIV/AIDS is multiple regression.

In multiple regressions we used SAS statistical software to see the effect of the continuous predictor variables on a continuous response variable which is behavioral risk (BHR).

We first check whether our data residuals are is approximately normal distributed when fitting the model. Also, we need to ensure that there is no serious multicollinearity among variables before we fit the model.

SAS Code 4.7

```
proc corr data=try;
var BHR AGE GEN ALCD NUMP RELB;
run;
```

The corresponding output follows:

Output 4.6

Pearson Correlation Coefficients

Prob > |r| under H0: Rho=0

Number of Observations

	BHR	AGE	GEN	ALCD	NUMP	RELB
BHR	1.00000	-0.04990	0.19029	-0.00904	0.31414	-0.05610
		0.4818	0.0068	0.8986	<.0001	0.4290
AGE	-0.04990	1.00000	-0.35534	-0.02472	0.02722	0.11329
	0.4818		<.0001	0.7276	0.7013	0.1093
GEN	0.19029	-0.35534	1.00000	-0.22167	-0.28228	0.17810
	0.0068	<.0001		0.0016	<.0001	0.0114
ALCD	-0.00904	-0.02472	-0.22167	1.00000	0.21901	-0.22056
	0.8986	0.7276	0.0016		0.0018	0.0017
NUMP	0.31414	0.02722	-0.28228	0.21901	1.00000	-0.24756
	<.0001	0.7013	<.0001	0.0018		0.0004
RELB	-0.05610	0.11329	0.17810	-0.22056	-0.24756	1.00000
	0.4290	0.1093	0.0114	0.0017	0.0004	

4.6.1 Analysis of Pearson Correlation Coefficients (output 4.6)

Output 4.6 displays the Pearson correlation coefficients between the variables in our model. We can see that our predictor variables are correlated with each other and with our response variable. The highest correlation is between BHR and NUMP with a correlation coefficient of 0.31414.

Moreover, co linearity diagnostics (shown in the appendix) on our variables have been carried out and these diagnostics are shown in the coefficient section of the output, under tolerance. All our variables have a tolerance above 0.7 meaning that there is no multicollinearity.
The corresponding table 4 shows multicollinearity diagnostics

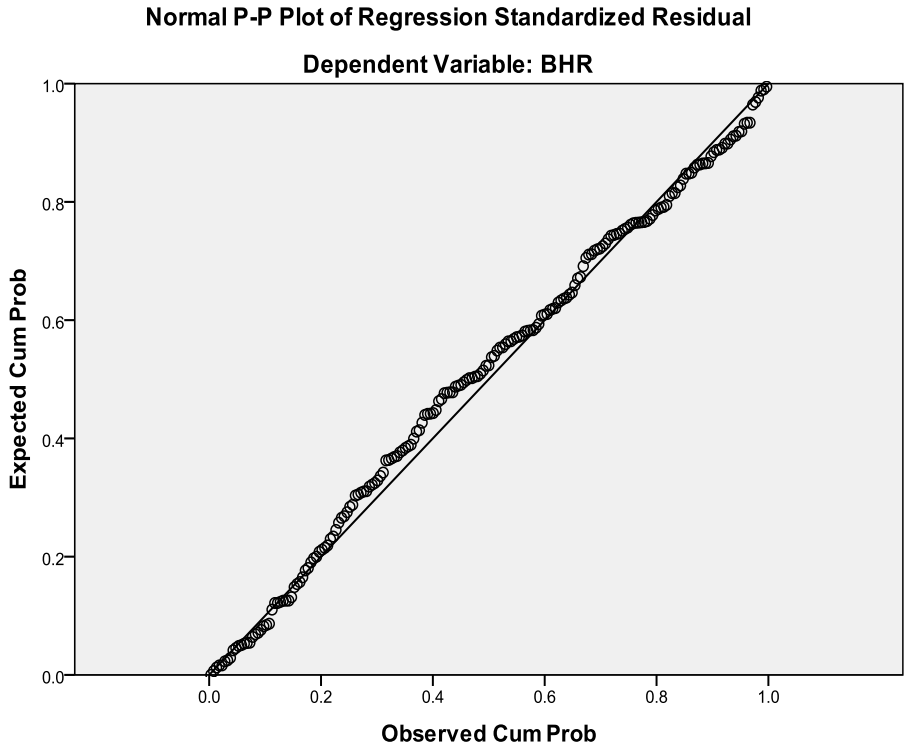
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	2.959	.381		7.757	.000		
	AGE	.013	.016	.056	.793	.429	.834	1.198
	GEN	.247	.057	.323	4.331	.000	.752	1.330
	ALCD	-.031	.082	-.026	-.382	.703	.896	1.117
	NUMP	.166	.029	.403	5.800	.000	.865	1.156
	RELB	-.015	.039	-.026	-.376	.707	.876	1.141

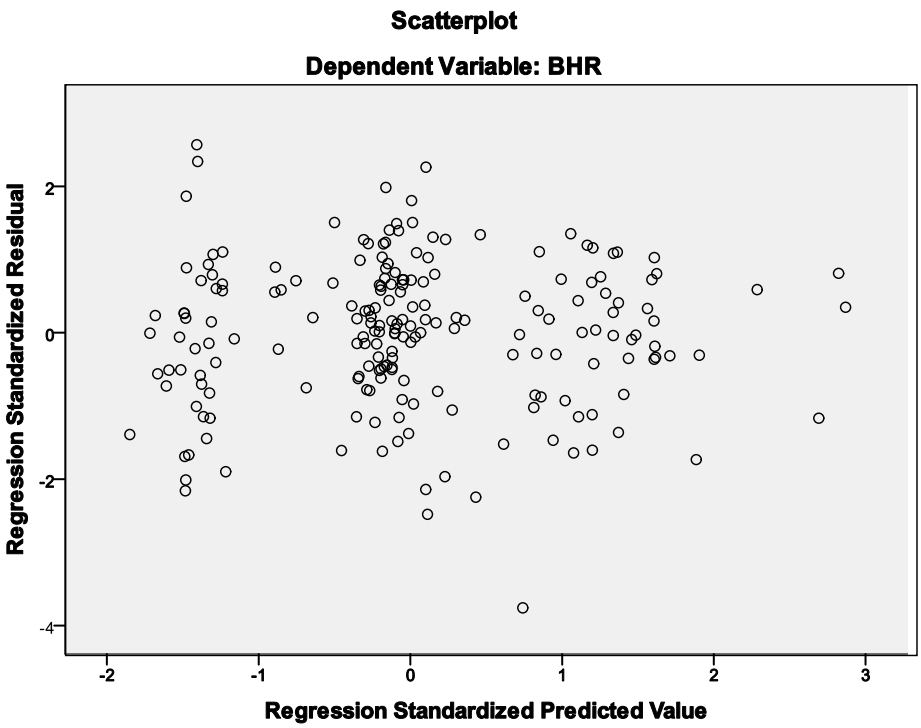
a. Dependent Variable: BHR

Furthermore, we can clearly see that the assumption of normality has not been violated too on the basis of the normal p-p plot and the scatter plot follows:

Figure 2: Normal P-P Plot of Regression



The corresponding Figure 3 is a scatter plot for multiple regression. The scatter plot provides a check for common variance of residuals over the range of response values.



After being satisfied that none of the underlying assumptions of multiple regressions were violated, we then fitted the multiple regression model using the SAS code below:

SAS Code 4.8

```
proc reg data=try;
model BHR=AGE GEN ALCD NUMP RELB;
RUN;
```

The corresponding output follows

Output 4.7 follows

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	5.50160	1.10032	8.97	<.0001
Error	196	23.90962	0.12261		
Corrected Total	201	29.41122			

Root MSE	0.35016	R-Square	0.1871
Dependent Mean	3.72488	Adj R-Sq	0.1662
Coeff Var	9.40064		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.96239	0.38105	7.77	<.0001
AGE	1	0.01250	0.01583	0.79	0.4306
GEN	1	0.24661	0.05700	4.33	<.0001
ALCD	1	-0.03572	0.07979	-0.45	0.6549
NUMP	1	0.16662	0.02868	5.81	<.0001
RELB	1	-0.01511	0.03917	-0.39	0.700

4.6.2 Analysis of Analysis of Variance and Parameter Estimates (output 4.7)

Adjusted R-square simply compensates for the number of explanatory variables in the model.

4.6.3 ANOVA

Examining at the model, we can see that the F-value is 8.97 and we have a p-value less than 0.0001. In Anova we are trying to determine how much of the variance is accounted for by our manipulation of the explanatory variables.

The ANOVA section helps us to assess the statistical significance of the chosen model terms. The F-statistic tests the null hypothesis that all the model parameters are zero. Our model attains statistical significance since the p-value is less than 0.05 and we infer at least one of the model coefficients is non-zero.

The R-square which tells us how much of the variance in the response variable is explained by the model. Here we obtain 0.1871.
We also have the Adjusted –R-square value in the output, which explains the variance in the response value if the sample size is small.

4.6.4 Parameter estimates.

Only the predictor variables GEN and NUMP attain low p-values, which imply that the two variables are making a strong unique contribution to explaining the response variables.

The other predictor variables are not statistically significant, and this outcome means that they are not making a unique contribution towards the response

variable in this data set. Their redundancy may be due to overlap with other explanatory variables in the model.

The parameter estimates can enable us to construct a multiple regression equation as follows.

$$Y = 2.962 + 0.013AGE + 0.247GEN - 0.036ALCD + 0.167NUMP - 0.015RELB$$

4.7 Conclusion

This chapter has applied three statistical methods to explore the effect of gender and age on the behavioral risk factors of HIV/AIDS. From the three statistical methods modeled, we note that the inferences were very similar, with the three models showing gender as the influential factor in determining one particular definition of the behavioral risk as the behavioral risk of HIV/AIDS.

However the predictor variable age, also a possible important factor in determining behavioral risk of HIV/AIDS was not a significant factor in this data set.

We have found that, besides gender, the number of partners has an association with the behavioral risk of HIV/AIDS.

Further discussions and conclusions will be made in chapter five.

CHAPTER FIVE

Conclusions and discussions

5.1. Introduction

The main objective of this study was to illustrate the statistical methods that can be used to show the influence of gender and age on the behavioral risk factors of HIV/AIDS. The study applied the logistic regression, loglinear regression and multiple regressions in fitting the data.

In section **5.3** we will make comparisons on the findings of the study and the results of similar studies.
In section **5.2** we will make discussions and conclusions In sections **5.4, 5.5** and **5.6** we will make discussions on recommendations, areas of future study and some concluding remarks respectively.

5.2. Discussions and Conclusions

The study sought to explore the influence of gender and age on risk behavior among students at one university. Gender and age were the main factors of interest, though they were associated with other factors like number of partners, religious beliefs and alcohol and drug abuse.

The study consisted of the response variable labeled as behavioral risk and defined by explanatory variables: gender, age, alcohol and drug and religious beliefs.

The logistic regression was the first statistical method applied to assess the influence of our main variables of interest. The response variable behavioral risk (BHR) was expressed as binary function with and was fitted against all five predictor variables. Actually, the method applied was the backward elimination process. The output obtained after using all five explanatory variables, showed only gender and number of partners as statistically significant variables.

The two predictor variables were very significant .This was because the interaction effects of the other three insignificant variables which were age, religious beliefs and drug abuse were eliminated on the basis of unacceptable p-values.

Further, we fitted a logistic model which consisted of our main variables of interest, age and gender .Only gender was significant
The study sought to investigate the significant associations amongst variables, when all variables were treated similarly as responses. We also included the variable number of partners as a factor in the loglinear model.

We wanted to examine the associations between the four variables namely behavioral risk (BHR), gender (GEN), age (AGE) and number of partners ($NUMP$).

The above saturated model gave a zero likelihood ratio (G^2) with the four way and three way interactions having very high $p-values$ thus showing their redundancy in a parsimonious model.
So we fitted a new loglinear model consisting of two way interactions only, and from the output we obtained a G^2 of 0.67 and non-significant $p-value\ of\ 0.9448$.
The likelihood ratio was too small so we decided to choose those only two way interactions which were significant.

Thus, we chose $BHR|GEN$, $AGE|GEN$ and $GEN|NUMP$ and we fitted them again. The results obtained gave a G^2 of 9.45 and a $p-value of\ 0.09$ which would be significant at a $10\%\ level$. So, dropping the two way interaction terms $BHR|AGE$ and $BHR|NUMP$ seemed to be useful as we managed to get better results as compared to the previous ones.

The non-significance of age on its own and the significance of gender exhibited by $p-values$ of 0.1211 and 0.001 respectively, confuse the similarity of the results of the loglinear and logistic regression models.

The third statistical method applied was multiple regression. To fit a valid multiple regression model we had to make sure that our explanatory variables are not severely correlated. In other words, we seek evidence of multicollinearity among variables before we fit the model.

Since in our study the main variables we were interested in were gender and age, it was very vital to notice that the highest negative correlation also occurred amongst the two variables.

We also explained the assumption of normality by fitting a P-P plot which showed that our data can be fitted linearly. A scatter plot was also made. The third method fitted a multiple regression model. The analysis of variance (anova) yielded had an $F-value of\ 8.97$ with a significant $p-value$. This value showed that the overall model has some explanatory value. The $R-Square\ of\ 0.1871$ denotes the fraction of the total squared error that is explained by the model.

The $Adjusted\ R\ Square$ which provides a better estimate for $R-Square$, when the number of predictor variables is large in relative to sample size n.

5.3. Comparison of the results of the study and the previous researches

We compare the results of three previous research studies with the results of our study. A study of gender power imbalance on women's capacity to negotiate self protection against HIV/AIDS in Botswana and South Africa, was carried out during the months of July and December 2003 by Tabitha T. Langen with the help of twelve research assistants in the Kwazulu Natal province of South Africa and Botswana, whereby a sample size of 2658 women aged 18-49 years were surveyed.

The statistical method used was logistic regression. A statistical model was fitted and the log likelihood (2546.998) from the full model containing all variables was compared with that of a simpler model containing only the gender power imbalance (3011.284) which resulted in a difference of 928.572. The results were highly significant with a p-value of 0.000. It was concluded that gender power imbalance significantly influenced women's ability to suggest condom use.

According to Langen this outcome rises b because women who did not openly suggest condom use to their partners may be afraid that increased partner communication about sexuality may disrupt power balance in intimate relationships, leading to marital conflicts, suspicions of cheating and even household partner violence. This influence can be supported by the fact that, from the number of women who suggested condom use 19 of them reported that they experienced violence from their intimate male partners upon suggesting condom use to them.

In our study we also used the logistic regression and the p-value of the variable gender was also significant which showed that gender played a significant role on the influence of the behavioral risk factors of HIV/AIDS.

A similar cross-sectional study of 385 drug users was carried out in 2006 the capital city of South Africa by Sarra L. Heden. In this study multiple regression and logistic regression methods were used.

A backward elimination process was used to assess gender interactions in terms of sexual risk behaviors, and a final model with a good fit using the Homer and Lemeshow Lack of Fit Test (chi-square (df)=8.59(8), P= 0.38) and a global null hypothesis test (chi-square(df)=88.44(23), P<0.0001) had the interaction terms such as gender and age included (Wald chi-square(df)=9.86(2),P<0.01).

This study also gave similar results with our own study in terms of gender but in their study both age and gender were significant in determining the behavioral risk factors of HIV/AIDS.

A third study evaluated the effectiveness of an HIV/AIDS prevention intervention for African American women at Virginia Commonwealth University. Participants were 398 heterosexual, unmarried African American women older than 18 years who agreed to participate in Sisters Informing Sisters on Topics about AIDS, a widely used HIV prevention program for African American women. The study was carried out by Faye Z. Belgrave and others. The logistic regression and multiple regression methods were also used.

It was hypothesized that as age increased, positive attitudes toward condoms would decrease. However, Age was not significantly associated with condom attitudes F (1, 289) = 0.11, p>0.05. One of the control variables, which was level of education had a small effect on condom attitudes F (3,290) = 2.87, p<0.05.

The second hypothesis predicted that age would be associated with HIV knowledge. It was hypothesized that as age increased, HIV knowledge would also decrease. It was found that age was not significantly associated with HIV knowledge F (1,281) = 0.26, p>0.05.

5.4. Concluding Remarks

In this study logistic regression, loglinear regression and multiple regressions gave us similar results .We can conclude that gender has got an effect on risk behavior of HIV/AIDS whilst age had a little effect if ever there was any.

As a result it is important that in future the emphasis must be on the influence of age. The main reason why much emphasis should be put on age is because; there are contradicting outputs on the effect of age from many previous studies which have been conducted.

Bibliography

[1] UNAIDS (2008): Report on the Global AIDS epidemic. Available at: http://www.unaids.org/en/KnowledgeCentre/HIVData/GlobalReport/2008/2008_Global_report.asp [Last accessed September 2010]

[2] Caldwell, J and Caldwell, P. (1999). Sexual regimes and sexual networking: The risk of an HIV/AIDS epidemic in Bangladesh.

[3] Buve, A Bishikwabo-Nsarhaza, K Mutangadura G. The spread and effect of HIV-1 infection in sub-Saharan Africa.

[4] FAO. (1997). Gender: the key to sustainability and food security. SD Dimensions, May (1997). Available at: http://www.fao.org/sd). [Last accessed September 2006]

[5] Roebuck J. When does old age begin? The evolution of the English definition. Journal of Social History. 1979; 12(3):416-28.

[6] McCullaghe, P and Nelder, JA. (1989). Second Edition: Generalized Linear Models

[7] Agresti, A. (1996). An introduction to Categorical Data Analysis. John Wiley & Sons: New York.

[8] Dobson, A. J. (2001). *An Introduction to Generalized Linear Models*, Second Edition. Chapman and Hall/CRC (November 2001), London.

[9] Dey, D. K., Ghosh, S. K., and Mallick, B. K. (eds.) (2000). *Generalized Linear Models: A Bayesian Perspectiv*e. Marcel Dekker, New York.

[10] Rice, J. C. (1994). "Logistic regression: An Introduction". B. Thompson, ed., *Advances in social science methodology*.

[11] Stadler J. Rumor, gossip and blame: implications for HIV/AIDS prevention in the South African Lowveld. AIDS Educ Prev.2003 Aug; 15(4):357-68.

[12] Maura E. Stokes, Charles S. Davis, Gary G. Koch: Categorical data analysis using the SAS system, second edition, (2005).

[13] Menard, Scott. (1995). Applied Logistic Regression Analysis

[14] Peng, Chao-Ying Joann; Lee, Kuk Lida; & Ingersoll, Gary M. (2002). An introduction to logistic regression analysis and reporting. Journal of Educational Research 96(1): 3-13.

[15] Strauss, David (1999). The many faces of logistic regression, American Statistician

[16] http://www.dtreg.com/logistic.htm [Last accessed September 2010]

[17] Afifi, A.A. Clark, Virginia University of California, Los Angeles, second edition 1990: Computer-aided multivariate analysis. Chapman & Hall

[18] Harman, (1968): Modern factor analysis, second edition. The university of Chicago press, USA.

[19] Sharma, Subhash. University of South Carolina, (1996). Applied Multivariate Techiniques.John Wiley & Sons: New York.

[20] http://biostatistics.oxfordjournals.org/content/10/2/327.short
[Last accessed September 2010]

[21] http://ncbi.nlm.nih.gov/pmc/artcicles/PMC286940
[Last accessed September 2010]

[22] http://www.graphpad.com/articles/Multicollinearity.htm
[Last accessed September 2010]

[23] Botswana National AIDS Coordinating Agency, author. Second Generation HIV/AIDS surveillance: *A technical Report*. Botswana: Gaborone; (2002).

[24] Population Reference Bureau . Demographic Data and Estimates for the countries and regions of the World.

[25] Coffin, J. M. (1999). Molecular biology of HIV. In The Evolution of HIV, ed. Crandall, K.A. 3-40.Baltimore: Johns Hopkins University Press.

[26] Goldstein, Matthew, W. R. Dillon (1978): Discrete discriminant analysis. John Wiley &Sons, Inc, New York.

[27] Menard, Scott (2002). *Applied logistic regression analysis, 2nd Edition*. Thousand Oaks, CA: Sage Publications. Series: Quantitative Applications in the Social Sciences, No. 106. First ed. (1995).

[28] Knoke, D, and Burke, P.J. (1980). Loglinear models. Sage publications, inc.newberry park, California, USA.

[29] Tabachnick, B.G.and L.S. Fidell (1996). Using multivariate statistics.3rd edition. Harper Collins. New York, USA.

[30] SAS Institute (1995). Logistic Regression Examples using the SAS System Version 6.

[31] UNGASS ((2010). 31st March) 'South Africa UNGASS Country Progress Report'

[32] Statistics South Africa, 'Mid-year population estimates July (2009).'

[33] Human Sciences Research Council (2009). 'South African National HIV Prevalence, Incidence, Behavior and Communication Survey, (2008). A Turning Tide among Teenagers?'

[34] http://en.scientificcommons.org/20386404
[Last accessed September 2010]

[35] http://www.springerlink.com/index/5672Q585VL377J44.pdf
[Last accessed September 2010]

[36] http://www.jbp.sagepub.com/cgi/rapidpdf/0095798409356686v1.pdf
[Last accessed September 2010]

[37] http://www.ncbi.nlm.nih.gov/sites/ppmc/articles/PMC1913066/.
[Last accessed September 2010]

Appendix A

Percent distribution of women who did not suggest condom use to their partners by back ground characteristics and area of residence.

	Botswana (N=1107)		KZN (N=1551)		Total (N=2658)	
Characteristics	% (n)	p-value[a]	% (n)	p-value	% (n)	p-value
Age difference between partners						
≤ 2 yrs	16.9(60)		25.5(155)		22.3(215)	
3–5 yrs	14.9(53)		34.1(182)		26.5(235)	
6–9 yrs	30.6(74)		27.2(76)		28.8(150)	
10+ yrs	60.3(94)	0.000	59.2(77)	0.000	59.8(171)	0.000
Woman's education						
≤ Primary	50.2(128)		61.3(111)		54.8(239)	
Secondary	19.2(142)		39.6(269)		28.9(411)	
Higher	10.1(11)	0.000	15.9(110)	0.000	15.1(121)	0.000
Marital status						
ever married	40.0(172)		45.4(229)		42.9(401)	
never married	16.1(109)	0.000	24.9(261)	0.000	21.5(370)	0.000
Abuse experienced						
all 3 types	29.8(67)		38.4(107)		34.5(174)	
2 types	30.9(94)		30.3(152)		30.6(246)	
Only 1 type	22.7(81)		37.8(185)		31.4(266)	
None	17.6(39)	0.001	16.4(46)	0.000	16.9(85)	0.000
Economically dependent						
Yes	34.0(185)		34.3(194)		34.2(379)	
No	17.1(96)	0.000	30.0(296)	0.078	25.3(392)	0.000
Aware of multiple partnerships						
Yes	27.9(161)		22.5(231)		24.4(392)	
No	22.7(120)	0.048	49.4(259)	0.000	36.0(379)	0.000
Total	**24.4(281)**		**31.6(490)**		**29.0(771)**	

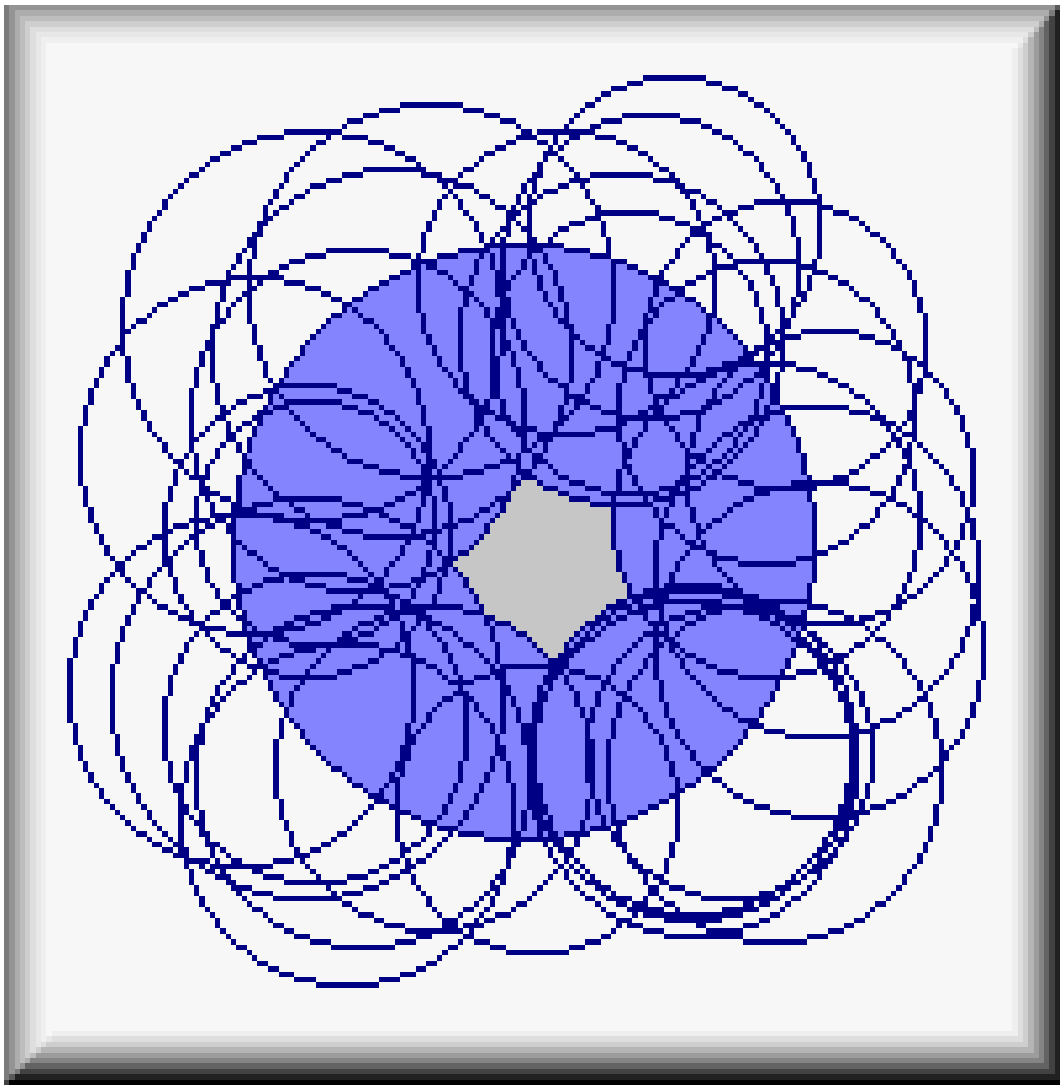
Appendix B

Percent distribution of women whose partners refused to use the condom by background characteristics and area of residence

	Botswana (N=829)		KZN (N=1059)		Total (N=1888) % (n)	
Characteristics	% (n)	p-value[a]	% (n)	p-value	% (n)	p-value
Age difference between partners						
≤ 2 yrs	15.5(46)		26.9(122)		22.4(168)	
3–5 yrs	16.5(50)		36.8(129)		27.4(179)	
6–9 yrs	26.8(45)		37.8(76)		32.8(121)	
10+ yrs	29.0(18)	0.003	54.7(29)	0.000	40.9(47)	0.000
Woman's education						
≤ Primary	33.1(43)		50.0(35)		39.0(78)	
Secondary	17.9(107)		39.6(162)		26.7(269)	
Higher	9.0(9)	0.000	27.4(159)	0.000	24.7(168)	0.000
Marital status						
ever married	35.2(92)		57.8(159)		46.8(251)	
never married	11.8(67)	0.000	25.1(197)	0.000	19.5(264)	0.000
Abuse experienced						
all 3 types	30.4(48)		49.4(85)		40.3(133)	
2 types	21.4(45)		37.8(131)		31.6(176)	
only 1 type	14.4(40)		26.9(82		20.9(122)	
None	14.2(26)	0.000	24.7(58)	0.000	20.1(84)	0.000
Economically dependent						
Yes	20.6(74)		34.1(126)		27.4(200)	
No	18.1(85)	0.378	33.3(230)	0.790	27.2(315)	0.903
Aware of multiple partnerships						
Yes	24.2(101)		36.0(286)		32.0(387)	
No	14.1(58)	0.000	26.4(70)	0.004	18.9(128)	0.000
Total	**19.2(159)**		**33.6(356)**		**27.3(515)**	

Appendix C

Venn diagram illustrating VIF



Appendix D

Informed consent form

Title

Statistical methods to model the influence of age and gender on the behavioral risk factors of HIV/AIDS.

Investigator(s) : Professor J.TYLER (040 602 2171)
: Professor Y.QIN (040 602 2162)
: Mr. B. Tlou (0727534672)

Purpose of the study

The main objective of the study is to explore the statistical methods that can be used to show the influence of gender and age on the behavioral risk factors of HIV/AIDS. The purpose of the study is to contribute to the monitoring of HIV/AIDS by selecting appropriate methods.

Description of the study

Participants will be provided with a questionnaire and they will be asked to respond to the questions in the questionnaire by choosing the preferred response. There is no actual or potential harm on the participants by completing the questionnaire. The only impact of the questionnaire will be consuming a 20 minutes of the participants' time. You may refuse to participate or may withdraw at any time.

Potential Benefits:

The HIV/AIDS research community will be the principal beneficiary and

Confidentiality:

Confidentiality will be respected and no information that discloses the identity of the participant will be released or published without consent unless required by law. This legal obligation includes a number of circumstances, infectious disease, and expression of suicidal ideas, where research documents are ordered to be produced by a court of law and where researchers are obliged to report to the appropriate authorities."

Participation:

Participation in research is voluntary. If you choose to participate in this study you may withdraw at any time.

Contact:
If you have any questions about this study, please contact:
Mr. Boikhutso Tlou
Room 5 Beda Ferguson
Alice 5700
University of Fort Hare
(0027)+27727534672
Email:btlou12@yahoo.com

Consent:

By signing this form, I agree that:

- The study has been explained to me.

  Yes **No**

- All my questions were answered.

  Yes **No**

- Possible harm and discomforts and possible benefits (if any) of this study have been explained to me.

  Yes **No**

- I understand that I have the right not to participate and the right to stop at any time.

  Yes **No**

- I understand that I may refuse to participate without consequence.

  Yes **No**

- I have a choice of not answering any specific questions.

  Yes **No**

- I am free now, and in the future, to ask any questions about the study.

Yes **No**

- I have been told that my personal information will be kept confidential.

Yes **No**

- I understand that no information that would identify me will be released or printed without asking me first.

Yes **No**

- I understand that I will receive a signed copy of this consent form.

Yes **No**

I hereby consent to participate in this study:

Name of Participant:

Signature:
Date:

Appendix E

Questionnaire

A. Background Information

1. How old are you.....

2. What is your gender? 1. Male 2.Female

3. What is your race?
    1. Black
    2. White
    3. Colored
    4. Asian
    5. Others (specify)……

4. Where do you live?
    1. On campus/hostel
    2. Off campus private accommodation
    3. Other

5. What is your marital status?
    1. Never married
    2. Married
    3. Separated
    4. Divorced
    5. Widowed

6. In which year of study are you now?
    1. First year
    2. Second year
    3. Third year
    4. Honors

7. What is your degree programme …?

B.Sexual Attitudes

8. How do you feel about having sex in the next 3 months?

1	2	3	4	5
VERY BAD	BAD	NOT SURE	GOOD	VERY GOOD

9. Would most people who are important to approve or disapprove you having sex in the next 3 months?

1	2	3	4	5
DISAPPROVE STRONGLY	DISAPPROVE	NOT SURE	APPROVE	APPROVE STRONGLY

10. Would your partner approve or disapprove of you having sex in the next 3 months? If you do not have a partner skip this question

1	2	3	4	5
DISAPPROVE STRONGLY	DISAPPROVE	NOT SURE	APPROVE	APPROVE STRONGLY

11. Would your mother approve or disapprove of you having sex in the next 3 months?

1	2	3	4	5
DISAPPROVE STRONGLY	DISAPPROVE	NOT SURE	APPROVE	APPROVE STRONGLY

12. Would your father approve or disapprove of you having sex in the next 3 months?

1	2	3	4	5
DISAPPROVE STRONGLY	DISAPPROVE	NOT SURE	APPROVE	APPROVE STRONGLY

13. Would your friends approve or disapprove of you having sex in the next 3 months?

1	2	3	4	5
DISAPPROVE STRONGLY	DISAPPROVE	NOT SURE	APPROVE	APPROVE STRONGLY

14. How do you feel about you using a condom if you have sex in the next 3 months?

1	2	3	4	5
VERY BAD	BAD	NOT SURE	GOOD	VERY GOOD

C. Sexual Behaviour

15. Have you ever had sexual intercourse (i.e., a man's penis inserted into your vagina)?

0. I have never had sexual intercourse
1. No
2. Yes

16. Have you ever been forced to have sexual intercourse against your will?

0. I have never had sexual intercourse
1. No
2. Yes

17. The first time you had sexual intercourse. Did your partner use a condom?

0. I have never had sexual intercourse
1. No
2. Yes

18. How old were you the first time you had sexual intercourse...?

19. How many different sexual partners do you have…?

20. Who influenced your sexual behavior most?

1. Mother
2. Father
3. Sister
4. Brother
5. Aunt
6. Uncle
7. Female cousin
8. Male cousin
9. Female friend
10. Male friend

21. The last time you had sexual intercourse, were you high on either alcohol or drugs?

0. I have never had sexual intercourse
1. No
2. Yes

22. The last time you had sexual intercourse, how many drinks did you have before having sexual intercourse?

0. I have never had sexual intercourse
1. (Write in)…….drinks

23. In the past 3 months, did you have sexual intercourse?

0. I have never had sexual intercourse
1. No
2. Yes

D. Alcohol and Drug Abuse

24. In the past month (30 days), on how many days did you smoke cigarettes?
(Write in)…….days

25. In the past month (30 days), on how many days did you drink alcohol?
(Write in)…….days

26. In the past month (30 days), on how many days did you have 5 or more drinks of alcohol?
(Write in)…….days

27. In the past month (30 days), on how many days did you use dagga (marijuana), mandrax?
(Write in)…….days

E. Number of Partners

28. Do you have a steady sexual partner?(If NO go to 31)

1. YES
2. NO

29. How many months have you and your steady partner been together? (Please answer in months)
……months

30. In the past 3 months did you have sexual intercourse with your steady partner?

1. YES
2. NO

31. In the past 3 months did you have sexual intercourse with someone who was not your steady partner?

1. YES
2. NO

32. In the past 3 months how often was a condom used when you had sexual intercourse with someone who was not your steady partner?

1. Never
2. Sometimes
3. Often
4. Almost every time
5. Every time

f. Religion

33. How often do you go to church or other religious activities?

1	2	3	4	5
Never	Few Times	About once a month	About once a week	Twice or more a week

34. How often do you read the bible, or other religious works?

1	2	3	4	5
Never	Few Times	About once a month	About once a week	Twice or more a week

35. How often do you listen to worship, church or gospel music?

1	2	3	4	5
Never	Few Times	About once a month	About once a week	Twice or more a week

36. How often do you listen to worship, church or gospel music?

1	2	3	4	5
Never	Few Times	About once a month	About once a week	Twice or more a week

37. How often do you listen to religious radio programmers?

1	2	3	4	5
Never	Few Times	About once a month	About once a week	Twice or more a week

38. How often do you say grace or pray before you eat?

1	2	3	4	5
Never	Few Times	About once a month	About once a week	Twice or more a week

39. How often do you pray before going to bed?

1	2	3	4	5
Never	Few Times	About once a month	About once a week	Twice or more a week

Ethical Clearance Form

OFFICE OF THE DEPUTY VICE-CHANCELLOR:
ACADEMIC AFFAIRS AND RESEARCH
Private Bag X1314, Alice 5700
Tel: 04060 22403
Fax: 0866282944
tsnyders@ufh.ac.za

University of Fort Hare
Together in Excellence

REC-270510-038

Application for clearance from the University of Fort Hare's Ethics Committee

Project title: **STATISTICAL METHODS WHICH CAN BE USED TO MODEL THE INFLUENCE OF AGE AND GENDER ON THE BEHAVIORAL RISK FACTORS OF HIV/AIDS**

Chief Researcher: Tlou Boikhutso

Supervisor: Professor Y Qin

Date of application: 13 April 2010

Having consulted the Dean of Research, I hereby grant permission to conduct the research.

Professor J R Midgley
Deputy Vice-Chancellor
Chairperson of the interim Ethics Committee

23 September 2010