IJRSET JANUARY 2021 Volume 8, Issue 1 Pages: 70-76

ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

RELATIONSHIP OF SEXUAL BEHAVIOR AND GENDER USING A LOGISTIC REGRESSION MODEL: (A CASE STUDY OF KISUMU COUNTY)

[1] Peter K. Kitsao,
[1] Statistician, School of Pure and Applied Science,
[1] University of Embu, Kenya.

ABSTRACT- Human immunodeficiency virus (HIV) represents a global, dynamic and unstable phenomenon. The nature of its occurrence in different regions in Kenya depends, among other determinants, on individuals and collective human behavior. The purpose of this study was to determine the sexual relations effect, the onset of first sex, their sexual partners and the use of condom used by these HIV/AIDS patients in the Kisumu. Secondary data of HIV/AIDS victims was used for this study. This data was analyzed using SPSS software to examine the sexual behavior and gender, using logistic regression analysis. The data was categorical and continuous in nature, where the predictor variables (Explanatory variables) are the sexual determinants and the gender being the dependent variable (Response variable).
The logistic regression model with six factors revealed some significant effects on the HIV/AIDS study. Thus, the determinants have a significant effect on gender. The logistic function revealed a significant association between the use of condoms, age first sex, Gender, marital status, why have STI and give receive for sex.

1. INTRODUCTION

Globally, HIV/AIDS havoc has caused panic since its infection has no cure. Over the years now, the phenomenon of the incurable disease has become prominent globally with serious implications of minimal survival of the victims. The presence of this disease in Kenya has transcended the level of worrisome of uncommon problem. Nationwide, especially in major cities, it is very common to find people living with the disease. Many researchers have been done that surround this study. In 2017, a research done by UNAIDS showed that 1.5 million people were living with the virus. Among those who were living with HIV in 2017, 110000 were children. The matter has been a public concern as well as a priority to the government and international organization (UNAIDS, 2018). A study done by *Lancet, 2009* indicated that the behavioral strategies could be an aim to reduce the virus widespread. For instance, the onset of the first intercourse, reduction of the number of sexual partners, use of condoms among other strategies(“August, 2009,” 2009). Therefore, sexual behavior change remains one of the prevention measures for further transmission of HIV/AIDS menace. Thus, this study seeks to investigate the sexual behavior of the victims of HIV virus in Kisumu County using a logistic regression to fit in it the model which is a predictive analysis. This will help

us to describe and explain the relationship between one dependent binary variable and independent variables.

2. STATEMENT OF THE PROBLEMS

Researchers have been done in the past highlighting the increased number of the epidemic which has wreaked havoc globally. There was need to explore the determinants that contribute to the increase rate of HIV/AIDS. Lancet, manuscript (2008) study shows that there are factors that are associated with the increased number of HIV victims such as the number of sexual partners, prostitutions among others. It is with significant effect that when one learns that he /she is HIV positive the knowledge becomes so profound in all his/her life. The aim of this project was to review the sexual behavior in relation with HIV transmission with the aid of logistic regression model. Since the dependent variable is binary in nature, it was necessary to use this kind of model to fit in the model.

It is clear that the use of logistic model is suitable in fitting a model since it explains the variable variation in the model. In this case, the model will help us describe the factors in terms of the amount of variance in the log odds. Thus, logistic regression model will enable us to determine the generalize ability of the model beyond this data on which the model is to be fitted.

3. LITERATURE REVIEW

Prior Researches have shown that logistic model is a popular tool in measuring the performance, developing a model and assessing and evaluating the goodness fit of abinary choice models. It has been applied widely in many situations and contexts involving dichotomous choice(Technology, 2012). In this research, the logistic model will enable us to explain the contribution of each explanatory variable in relation to the Response variable (Gender), minimizing the influence of the other independent variables. Since this Research is done in a Health setting, then it is important to consider the four main multivariate methods used in health science to analyze the data which are linear Regression, Logistic Regression, Discriminate analysis and proportional hazard Regression(Ae, 2013). Both Logistic model and discriminate analysis are methods for analyzing categorical-response variables. However, many statisticians prefer Logistic regression model to Discriminate analysis because Logistic Regression analysis does not assume the predictor variables as normally distributed, like discriminate analysis does(Statistical & Ncss, n.d.). Since we wish to report the performance of this model as well as the goodness of fit and the development of the model, then binary logistic regression is the most versatile. It can perform predictor variable subset selection search to determine the best regression model with the least number of predictor variables. It also provides confidence intervals on the predicted values, and ROC curves which aid in determining the sensitivity and specificity cutoff point of the binary classifier(Statistical & Ncss, n.d.).

In this research and other related studies, the participants targeted are people living with HIV/AIDS who are part and parcel of the community. The predictors considered in this research are use of condoms, age first sex, and marital status, why have STI and give receive for sex. Furthermore, logistic regression analysis will predict the mean value of gender from the known values of these independent variables (Topic, 2000). Binary logistic regression analysis allows you to validate your results by automatically classifying rows that are not in use during the analysis(Statistical & Ncss, n.d.).This project was meant to explore and utilize the knowledge of logistic regression model to recommend the way of living of the people with HIV virus in our society. The assessment and evaluation of the performance of the study is for the purpose of validation of the model(Assessment, 2003). Moreover, evaluation of a model based on classification problems using confusion matrix and ROC curve can be used to improve a

logistic regression model ((Yang & Berdine, 2017). It is therefore prudent to have this study which will highlight some important factors that when put into consideration, will benefit the entire community and improve the lives of the patients living with disease.

4. RESEARCH MATERIAL AND METHODOLOGY

Here, we will highlight the logic in which the method used was selected for this research, cases and the statistical tools used. The Chapter will farther provide the guidelines for the purpose of applying Logistic Regression Model. Therefore, this chapter will cover the research processes, Research design, the data collection procedures and recording and analysis of the data.

Research design

This research used a sample for 2009 STIs data that represents the population of the entire county. The probability sample representative of the collected data of almost 219 based on assumption was sampled out for the people living with HIV virus. The method used for collection of the data was sampling process since it permits every element (HIV patient) in the sampling frame (population with the disease) to have an equal chance. This project was a case study design and it was meant to make the researcher to understand the behavior of the target population (HIV patients)being studied through logistic model. Logistic regression analysis is one of the types of multivariate analysis techniques which is very popular in research institutions. Logistic regression model enables organizations to create knowledge and thereby improving their decision making. In this case, Logistic regression model will allow the project to understand the operations of the people living with HIV in real life situation.

Statistical Methods

The obtained data for 2009 STIs sample size of 219 was used to represent the population living with HIV virus in Kisumu County. The sample obtained contained continuous and categorical variables which enhanced the use of logistic regression analysis to be suitable for this research. The statistical significance of the individual regression coefficients is evaluated by the Wald Chi-square statistic (Peng & Ingersoll, 2014) . Even though Logistic regression analysis competes with discriminate analysis as a method for analyzing categorical-response variables, but still logistic regression is more versatile and better suited for modeling most situations than is discriminate analysis. This is because the logistic regression does not assume the independent variables to be distributed normally as Discriminate does(Statistical & Ncss, n.d.). Some of the statistical processes the data went through are discussed below.

Data Collection

The obtained data for 2009 STIs sample size of 219 was collected from secondary sources. The characteristics under consideration (study) are listed below;

- Age
- Occupation
- Marital status
- Educational level
- Religion
- Weight and Height
- Duration of illness
- Who infected you
- Take Alcohol
- Use of condom
- Type of STIs
- Relationship
- Gender
- Onset of first sex
- Sexual partners

Data Analysis

Analysis of data was done to the already cleansed data, that is, removal of the irrelevant parts of the data, correcting and finally

modifying the data. The data was copied from Excel sheet and pasted to SPSS. To analyze continuous and categorical variables Binary logistic regression analysis was used to generate the Classification tables.

Logistic Regression Model

Logistic regression is a technique used to make predictions when the predicted variable is a binary (coded as 0 and 1) and the predictors are continuous and/ or discrete(Hosmer, D. W. & Lemeshow, S. 1989).Logistic regression models are suitable for building a biological acceptable model that can identify the relationship between the response variable and the predictor(s) in a way that is best suitable for the use of the least variables(Dergisi & Korkmaz, 2012).It assumes that the relationship between the predicted variable and the predictor(s) are logarithmic.

It allows one to predict a discrete output such as a group membership from a set of variables(Atitwa, 2005).This method is an alternative to the linear regression model when the normality assumption fails in case of binary, categorical or multi-categorical discrete variables(Dergisi &Korkmaz, 2012).Since the model for analysis is mathematically flexible and interpretation is simpler, hence the increased use of the method. Furthermore, logistic regression has minimal assumption limitation (Özdamar, 2002).In logistic regression, mathematical model of a set of predictor variables is used to determine a logit transformation of the predicted variable. Since the numerical values of 0 and 1 are assigned to the two outcomes of a binary variables. Often, the 0 indicates a negative response and the 1 represents a positive response. The mean of this variable will be the proportion of positive responses. If pis the proportion of observations with an outcome of 1, then 1-p is the probability of an outcome of 0. The ratio p/(1-p*)* is called the odds and the logit is the logarithm of the odds, or just log odds. Mathematically, the logit transformation is written;

l= logit(p)=ln(p/1-p)

For a single exposure variable Y, the model takes the form

$$\ln (p/1-p) = a+bx \quad (1)$$

Where p denotes the probability of occurrence of the outcome D and x is the value of an exposure E. The equation can be inverted to give an expression for the probability of p as;

$$p(D) = \frac{1}{1+\exp(-a-bx)} \quad (2)$$

The risk of the outcome given the exposure wills thus be obtained by putting x=1 in the equation (2), we obtain,

$$p(D/E) = \frac{1}{1+\exp(-a-b)} \quad (3)$$

while the risk of the outcome given no exposure (x=0) we obtain,

$$p(D/E) = \frac{1}{1+\exp(-a)} \quad (4)$$

The relative risk is the ratio of these two expressions. We will use the odds and odds ratio. The odds of the outcome given exposure are, from equation (3),

$$\frac{p(D/E)}{p(D/E)} = \frac{p(\frac{D}{E})}{1-p(\frac{D}{E})} = \frac{\frac{1}{1+\exp(-a-b)}}{1-\frac{1}{1+\exp(-a-b)}} \quad (5)$$

Which reduces to exp (a + b).Finally obtain the odds Ratio (OR) as

$$OR = \frac{\exp(a+b)}{\exp(a)} = \exp(b) \quad (6)$$

This means that the parameter bin the model is the natural logarithm of the odd Ratio.

Multiple Binary Logistic Regression Model

If there are p predictor variables x_1, x_2,...,x_p, the general form of multiple logistic regression model is as follows;

$$p(D) = \frac{1}{1+\exp(-a-\sum_{j=1}^{p} b_j x_j)} \quad (7)$$

Parameters b_1,...,b_p, were estimated using the maximum likelihood method. The parameter should give the significance of each independent variable to the outcome D. The estimated parameter forming the model was used to classify the remaining part of the data b. The cut value is .500 a. Constant is included in the model. into either of the two groups.

EXPLORATION OF THE RESULTS

The results of the Logistic Regression Model Prediction of the dependent variable

Observed		Predicted		
		Gen		Percentage Correct
		Female	Male	
Gen	Female	0	105	.0
	Male	0	114	100.0
Overall Percentage				52.1

Table1 Classification Table[a,b]

The constant (intercept) of the model is only represented by the Block 0 output. To calculate the base rates of the two decision options; 114/219= 52% predicted to be male, 48% predicted to be female. The best strategy is to predict for every case that the subject gender is male. Using this strategy, then the 52% of the time is correct.

		B	S.E.	Wald	Df	Sig.	Exp(B)
Step 0	Constant	.082	.135	.370	1	.543	1.086

The table 2 above shows the intercept-only model is ln(odds)= 0.082. The exponential of the predicted odds {Exp(B)}= 1.086. Since the predicted odds for gender female is 1.086. For 114 of our subjects predicted to be gender male and 105 predicted to be gender female. Therefore, the Observed odds are 114/105= 1.086.

Classification of the independent variable

		Chi-square	df	Sig.
Step 1	Step	30.590	4	.000
	Block	30.590	4	.000
	Model	30.590	4	.000

Table 3 Omnibus Tests of Model Coefficients

Block 1 output is the predictor variables. The Omnibus Tests of Model Coefficients gives a Chi-Square of 30.590 on 4 df, significant beyond 0.001.This is a test of the null hypothesis that adding up the predictor variables to the model has not significantly increased the ability to predict the decisions made by the subjects.

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	272.639[a]	.130	.174

Table 4 Model Summary

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

The above table shows that the -2 Log Likelihood statistic is 272.639. This statistic measures how poorly the model predicts the prediction. The smaller the statistic the better the model. The Cox and Snell R- Square (0.130), but cannot reach a maximum value of 1. The Nagelkerke R- square (0.174) can reach a maximum of 1.

	Observed		Predicted		
			Gen		Percentage Correct
			Female	Male	
Step 1	Gen	Female	83	22	79.0
		Male	56	58	50.9
	Overall Percentage				64.4

Table 5 Measuring the performance of the model

a. The cut value is .500

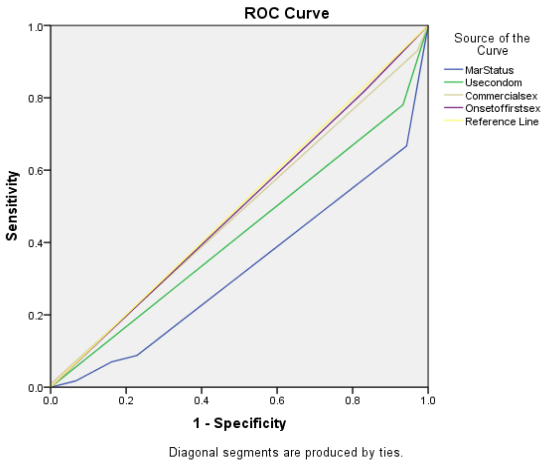
The Classification Table allows us to correctly classify 58/114= 51% of the subjects where the predicted gender male was observed. This is the Sensitivity of prediction.

P= correct⁄event did occur

This rule also allows the correct classification of the subjects where the predicted event was not observed. This is known as the sensitivity of prediction which is the percentage of the non-occurrences correctly predicted.

P= correct⁄event did not occur =83/105 =79%

The overall predictions that were correct 141 out of 219 cases and the overall success rate was 64% which is an increase of the intercept only that was 52% for the model.



Since Marstatus and Use condom are far from the reference line, thus shows that the model is good. An overlapping curve such as the case of Onsetoffirstsex and Commercial sex will result to bad model.

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	MarStatus	-.742	.192	14.966	1	.000	.476	.327	.693
	Commercialsex	-1.081	.740	2.135	1	.144	.339	.080	1.446
	Usecondom	-1.163	.494	5.547	1	.019	.312	.119	.823
	Onsetoffirstsex	-.288	.382	.567	1	.451	.750	.355	1.586
	Constant	4.214	1.209	12.139	1	.000	67.616		

Table 6 Variables in the Equation

The Table above shows the logistic regression coefficient, Wald test and odds ratio of each of the predictors. Using 0.05 level of significance; Marstatus and use condom had a significant effect on predicting the gender of the model. However, Commercial sex workers and onsetoffirstsex had no significant effect on predicting the gender. Since out of the predictors only two of the predictor variables are less than 0.05 (Marstatus 0.000<0.05, and Use condom 0.019<0.05, and the constant 0.000<0.05) and the rest are not statistically significant. Therefore, not included in the predicting model. The Wald statistic test gives the unique contribution of each predictor variable (Marstatus 0.192, Commercial sex 0.740, Use condom 0.494, Onsetoffirstsex 0.382). Below is the generated prediction equation from the Table.

ln(odds)= 4.214 -0.742Marstatus -1.163Usecondom

The table 6 proves that there is a relationship between Marstatus, Use condom and Gender. The general model level of significance is below 0.05 hence the model is statistically significant.

CONCLUSION

This project study was mainly focused on sexual practices of HIV/AIDS victims in the Kisumu County. It was defined by; measuring the performance of the Logistic regression model and developing the model that can predict the sexual practices of HIV victims. The analysis was done by Binary Logistic regression to determine the impact of sexual behavior on gender. The main factors in question were; Marstatus, Use condom, Commercial sex and Onsetoffirstsex. The significant effect of each independent variable was observed on the binary variable and an inference was arrived. The model was statistically significant between gender and use condom and Marstatus.

REFERENCE

[1]. Ae, H. (2013). An Introduction to Logistic Regression : From Basic Concepts to

Interpretation with Particular Attention to Nursing Domain. 43(2), 154–164.Assessment, G. (2003). IN LOGISTIC REGRESSION.
[2]. Atitwa, E. B. (2005). Socio-Economic Determinants of Low Birth Weight in Kenya : An Application of Logistic Regression Model. 4(6), 438–445. https://doi.org/10.11648/j.ajtas.20150406.14
August, 2009. (2009). In The Lancet Neurology. https://doi.org/10.1016/S1474-4422(09)70284-4
[3]. Dergisi, f., & korkmaz, m. (2012). The importance of logistic regression implementations in the turkish livestock sector and logistic regression implementations / fields. 16(2), 25–36.
[4]. Peng, C. J., & Ingersoll, G. M. (2014). An Introduction to Logistic Regression Analysis and Reporting. September 2002. https://doi.org/10.1080/00220670209598786
Statistical, N., & Ncss, S. (n.d.). Logistic Regression. 1–69.
[5]. Technology, C. (2012). Building and evaluating logistic regression models for explaining the choice to adopt MOOCs in India Sangeeta Trehan and Rakesh Mohan Joshi Indian Institute of Foreign Trade , India. 14(1), 33–51.
[6]. Topic, D. (2000). Literature Review ●. 626. UNAIDS. (2018). 2017 GLOBAL HIV STATISTICS. In Ending the AIDS Epidemic.
[7]. Yang, S., & Berdine, G. (2017). The receiver operating characteristic (ROC) curve. 5(19), 34–36. https://doi.org/10.12746/swrccc.v5i19.391.