ESTIMATION OF HIV PREVALENCE AMONG WOMEN IN KENYA IN THE PRESENCE OF MEDIATION USING LATENT TRAIT ANALYSIS

1

Abstract

Estimating prevalence in cause-effect relationships where the mediator variables are assumed to be latent is not usually easy. However, the use of proper indicators and statistical model can make the measurement and use of such constructs easy. Structural Equation Modeling makes it possible to analyze simultaneously both the relationship between the latent variable and the links between the latent variable and their indicators.The 2018 Kenya AIDS Indicator Survey data was used to validate the model developed. The maximum likelihood was used to estimate the model parameters. The findings of the study were, there is a relationship between education attainment and knowledge /awareness of HIV/AIDS. The results further show that education level is not associated with HIV prevalence after controlling for a number of socio-demographic characteristics and behavioral factors. These findings can inform policy makers in formulation of appropriate HIV/AIDS management (policies) and intervention strategies aimed at reducing HIV/AIDS prevalence that has remained a challenge in many developing countries.

Subject Classification: xxxxxx

Keywords: HIV prevalence, continuous latent mediator, latent trait analysis, observed indicators

1 Introduction

According to Kharsany[1] HIV/AIDS has been here for more than thirty years. To date, there is still no cure nor an effective vaccine for the disease. However, the inevitably fatal

disease has been transformed to a chronic and manageable condition leading to substantive decline in the worldwide rates of AIDS related deaths and new infections. This is due to intervention strategies such as; introduction of anti-retro-viral therapy (ART), focusing on high transmission areas and key populations and by implementation of evidence-based prevention strategies like voluntary male circumcision, prevention of mother to child transmission among others. Despite the prevention efforts above, death rates still remain high since a large number of people remain unaware of their HIV status and therefore fail to be adequately linked to care and treatment programs, [**?**, 3]. In 2018, an estimated 35.0 million people were living with HIV worldwide where Sub-Saharan Africa (SSA) accounts for 71% of the population yet is home to only 12% of the global population.
Transmission of HIV in SSA is majorly through sexual contact and has remained a challenge to the possibility of an AIDS free generation, despite there being many initiatives to prevent it. Globally, 15% of women living with HIV are aged $15-24$ years, of whom 80% live in SSA, [4]. The NASCOP report suggests that, prevention efforts must focus on broad social movements that contribute to safer sex behaviors in young women and extend to the general population with increased vulnerability to HIV. The report further explains that, the goal of HIV intervention research is to develop interventions that encourage participants to reduce or eliminate sexual or social behaviors that put themselves or others at increased risk for HIV infection.
The study sought to establish how the education levels of Kenyan women aged between 15-64 years affect their HIV status through unobserved sexual behaviour patterns measured by three categorical indicators (use of condoms, non-marital sexual relationship and abstinence). The surveyed data were measured on different metric scales, that is; binary outcome (HIV status), continuous latent mediator (sexual behaviour patterns), and categorical predictor variable (education levels). Such a dynamic and complex multivariate model calls for a powerful statistical analysis tool/ technique. Structural Equation Modeling (SEM), and more specifically, the latent trait analysis (LTA) was used.

2 Literature Review

Many scholars are interested in understanding the process by which an independent variable affects a dependent variable either directly or indirectly through a mediator. Researchers have tested for mediation when all the variables are continuous, but a definitive answer had been lacking as to how to analyze the data when predictor variables are categorical, latent mediator variables are continuous and with categorical indicators and a binary outcome. Survey data in economics, social and medical fields contain variables which are measured on different scales. They may be on binary scale, categorical scale (nominal or ordinal), metric scale (discrete or continuous) or a combination of the above. Researchers have proposed various solutions to working with categorical variables or a mix of categorical and continuous variables in mediation. For example;[5] Hayes, Preacher and Myers focused on predictor X and allowed it to be categorical, not just binary and sug-

gested that it would be ideal to allow for categorical M and Y as well. Other researchers make the assumption that, while a manifest variable may be discrete, the underlying construct is continuous. For example,[6] and [7] proposed modifications to structural equation modeling, of which the mediation form is a special case, for categorical variables. Muthen assumed categorical predictor variables, continuous latent variable with observed categorical data arising through a threshold step function. His estimation procedure was based on generalized least squares, assumed normal distributions, and recommended a large sample size. Motivated by Muthen's study, the current study will use latent trait analysis to test for mediation and report findings when predictor variables are categorical, latent mediator variables are continuous with categorical indicators and a binary outcome.
In a study on substance use as a mediator of the relationship between life stress and sexual risk among young transgender women, Anna et al [8] used mediation analysis to determine whether life stress through the mediator "substance use" was associated with increased sexual risk among young trans-gender women in South America. The analysis was based on data collected from 116 trans-gender women aged 16-25 years as part of the baseline assessment for an HIV prevention intervention. The median age was 20 years. Controlling for age, high life stress was associated with an increased odds of sexual risk. This association was attenuated when substance use was added to the model. The results indicated a statistically significant indirect effect. They hypothesized that one of the pathways through which stress impacts sexual risk is illicit substance use, that is, exposure to stress is associated with elevated levels of substance use, which in turn lead to greater sexual risk taking as discussed by Dohrenwend [9].
Bryan et al[10] used path analysis to examine the effects of a condom promotion intervention on condom use intentions in a sample of college women, as mediated by theory-based model variables. The authors developed a theoretical model of condom use specifically tailored to the young women. It was hypothesized that, the intervention would change perceptions of sexuality for women perceived to be susceptible to common sexually transmitted diseases (STDs) and that it will bring out benefits of condom use. Although the final model indicated that the intervention had a direct impact on the perceptions of sexuality for women and an indirect effect on benefits of condom use, it didn't assess the reliability and validity of the observed indicators. These two studies had shortcomings in the hypotheses because of the use of manifest variables. Our study intends to address this by the use of latent variables as opposed to manifest variables since latent variable approach to mediation is more powerful than approaches with only manifest variables such as simple path or regression analysis as discussed by Aminger, Fritz et al[11, 12]..

3 Methods

Structural equation modeling (SEM) is a multivariate statistical technique that allow the examination of a set of relationships between both independent and dependent variables, which are either continuous/discrete or observed/unobserved [13, 14]. Both of these vari-

ables can either be factors (latent) or measured (observed/ manifest). SEM seeks to represent hypothesis about the means, variances and covariance of observed data in terms of a smaller number of "structural" parameters defined by a hypothesized underlying theoretical model as shown by Bentler [15]. It is an essential tool in statistical analysis in scientific research due to its flexibility. A major feature in the development of SEM is to conceptualize latent variables from the observed (indicators) variables given. Latent variable models refer to structural equation models that take measurement error into account when statistically analyzing data.
A mediator (also called Mediating or intervening or intermediate variable) is a variable that is between the independent and dependent variable(s)as shown by Mackinnon5 [16]. They explain how and why an intermediate variable transmit the effect of independent variable on an outcome (dependent variable) as illustrated by Mackinnon3 [17]. Mediators are also called process variables since they are variables that describe the process by which an independent variable affects a dependent variable. This was discussed by Judd [18]. A mediation model can also be explained diagrammatically using the path diagrams as illustrated in Figure 1. In the Figure 1(a), the effect of X on Y, represented by c is called X's total effect on Y. This total effect is interpreted as the expected change in Y when X is changed by one unit. The effect of X on Y might come directly or indirectly. Figure 1(b) represent the simplest mediation model, in this model, variable X has a direct effect on Y, denoted as c'. The variable X is also hypothesized to affect the mediator M, which then has an effect on variable Y. The effect of variable X on variable Y through variable M is called the indirect effect or mediated effect which was discussed by Angela, Baron, Kenny [19, 20, 21]. The direct and indirect effects are respectively given by the Equations 1 and 3.

$$Y = c_0 + cX + e_y \tag{1}$$

$$M = a_0 + aX + e_m \tag{2}$$

$$Y = d_0 + bM + c'X + e_y \tag{3}$$

3.1 Model Formulation

The latent trait model is divided into two

(i) Structural Model

(ii) Measurement Model

3.1.1 Structural Model

The structural model equation is further subdivided into two i.e.

$$M = \beta_0 + \beta x + \varepsilon_1 \tag{4}$$

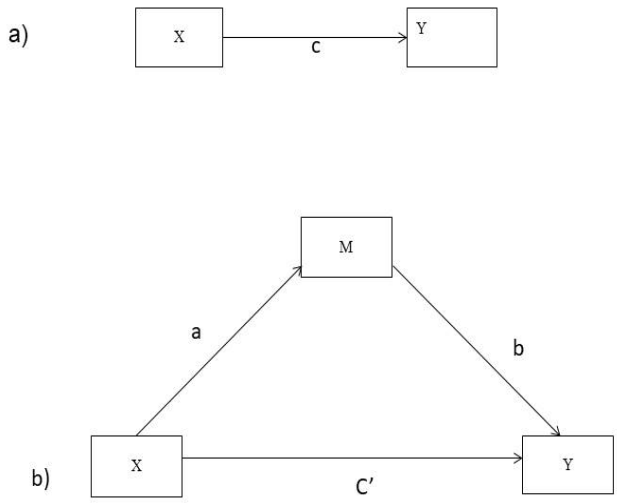

Figure 1: Path diagram

and

$$Y = \gamma_0 + \gamma_1 m + \gamma_2 x + \varepsilon_2 \tag{5}$$

Equation 4 and Equation 5 can now be written in matrix form as is as given below;

$$\begin{bmatrix} M \\ Y \end{bmatrix} = \begin{bmatrix} \beta_0 & 0 & 0 \\ \gamma_0 & \gamma_1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ m \\ y \end{bmatrix} + \begin{bmatrix} \beta_1 \\ \gamma_2 \end{bmatrix} x + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

Where m is a mediator variable, y the outcome variable while ϵ_1 and ϵ_2 ,are random errors with expected values (means) of zero and are uncorrelated with the x. A constant is absent from the equations because the variables are deviated from their means. This deviation form will simplify algebraic manipulations but does not affect the generality of the analysis. γ_2 is the structural parameter that indicates the change in the expected value of y after a one-unit increase in m holding ϵ_2 constant. γ_1 and γ_2 are regression coefficients β_1 is the latent coefficient. The matrix can then be compactly written as

$$\eta = \beta\eta + \Gamma\mathbf{x} + \zeta \tag{6}$$

Equation 6 is the structural model equation where η is a vector representing dependent variables (M and Y), X is a vector representing independent variables (X) and ζ is a vector representing random errors ϵ_1 and ϵ_2 respectively. The parameters β and Γ are as defined above.

3.1.2 Measurement Model

The measurement model gives the relationship between observed dichotomous variables and unobserved latent variable. It describes how each latent variable is defined via the manifest variables, and provides information about the validity and reliability of the structural model. The measurement model is given by;

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \gamma_{01} & \gamma_{11} & 0 \\ \gamma_{02} & \gamma_{12} & 0 \\ \vdots & \vdots & \vdots \\ \gamma_{0n} & \gamma_{1n} & 0 \end{bmatrix} \begin{bmatrix} 1 \\ m \\ y \end{bmatrix} + \begin{bmatrix} \gamma_{21} \\ \gamma_{21} \\ \vdots \\ \gamma_{2n} \end{bmatrix} x + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

which can be simplified to;

$$Y = \gamma_1 \eta + \gamma_2 \mathbf{x} + \varepsilon \tag{7}$$

From the structural equation 6, we make η the subject of the formula as follows

$$\eta - \beta\eta = \Gamma\mathbf{x} + \zeta \tag{8}$$

$$\eta(\mathrm{I} - \beta) = \Gamma\mathbf{x} + \zeta \tag{9}$$

$$\eta = (\mathrm{I} - \beta)^{-1}\Gamma\mathbf{x} + (\mathrm{I} - \beta)^{-1}\zeta \tag{10}$$

Substituting Equation 10 into equation 7, we have the following equation

$$y = \gamma_1[(\mathrm{I} - \beta)^{-1}\Gamma\mathbf{x} + (\mathrm{I} - \beta)^{-1}\zeta] + \gamma_2\mathbf{x} + \varepsilon \tag{11}$$

which gives the general formula of the measurement model as below

$$y = \Lambda_y \mathbf{x} + \varepsilon \tag{12}$$

where, $\Lambda = (\beta, \eta, \Gamma, \gamma_1, \gamma_2, \sigma_\eta, \sigma_y)$, y is a vector of the observed endogenous variables and ε is a vector of errors. Λ_y is the matrix of the structural coefficients between the observed variables and the latent variables.

3.2 Direct and Indirect Effect

.

In Figure 2, we illustrate the direct and indirect effects of education levels to HIV outcome. The direct effect is that influence of one variable on another that is unmediated by any other variables in a path model represented by Figure 2a. The indirect effects of a variable are mediated by at least one intervening variable shown by Figure 2b.

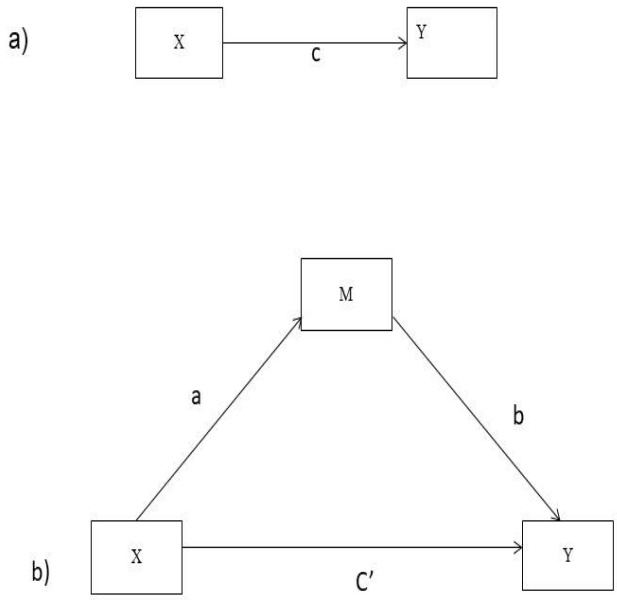

Figure 2: Path diagram

3.2.1 Estimation of Direct Effect

Figure 2a represents the first model which is a direct effect model, it assumes that the only effect on the outcome Y is a direct effect from exposure on predictor X. i.e, the direct effect model assumes that there is no mediation occurring between X and Y. As such, we can assume Y to be a linear function of its predictor X and it is assumed that the errors are normally and independently distributed with constant variance . In the measurement model denoted by Equation 12, the normal density function for the errors is

$$f(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} exp[\frac{1}{-2\sigma^2}\varepsilon_i^2] \qquad\qquad i = 1, 2, 3....n \tag{13}$$

The likelihood function is the joint density of $\varepsilon_1, \varepsilon_2 ..... \varepsilon_n$ given as

$$L(\beta, \sigma^2) = \prod_{i=1...n} f(\varepsilon_i) \tag{14}$$

$$= (\frac{1}{2\pi\sigma^2})^{n/2} \exp[-\frac{1}{2\sigma^2} \sum_i \varepsilon^2] \tag{15}$$

$$= (\frac{1}{2\pi\sigma^2})^{n/2} \exp[-\frac{1}{2\sigma^2} \varepsilon'\varepsilon] \tag{16}$$

replacing the value of ε

$$= (\frac{1}{2\pi\sigma^2})^{n/2} \exp[-\frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta)] \tag{17}$$

Since the log transformation is monotonic, so we maximize ln L(β , σ^2) instead of L(β , σ^2) to get

$$\ln L(\beta, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2)(y - X\beta)'(y - X\beta) \tag{18}$$

The maximum likelihood estimators (mle) of β and σ^2 are obtained by equating the first order derivatives of ln L (β , σ^2 ,) with respect to β and σ^2 to zero as follows:

$$\frac{\partial \ln L(\beta, \sigma^2)}{\partial \beta} = \frac{1}{2\sigma^2} 2X'(y - X\beta) = 0 \tag{19}$$

upon simplification, we get

$$\frac{\partial \ln L(\beta, \sigma^2)}{\partial \beta} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta) \tag{20}$$

The likelihood equations are given by

$$X'X\beta = X'y \tag{21}$$

$$\sigma^2 = \frac{1}{n} (y - X\beta)'(y - X\beta) \tag{22}$$

Since rank(X) = k so that the unique mle of β and σ^2 are obtained as

$$\hat{\beta} = (X'X)^{-1} X'y \tag{23}$$

$$\hat{\sigma^2} = \frac{1}{n} (y - X\hat{\beta})'(y - X\hat{\beta}) \tag{24}$$

3.2.2 Estimation of Indirect Effect

The indirect effects of a variable are mediated by at least one intervening variable. The sum of the direct and indirect effects is the total effects. In our model the indirect effect is illustrated by the influence of X on Y through the intervening variable η. We expect a one unit change in X to lead to an expected β change in η. Thus β change in η should lead to an expected γ_1 change in Y. Therefore the indirect of X on Y is β γ_1. Figure 2b represents an indirect as well as direct effects. [19] The model assumes not just a direct

effect from the predictor X to the outcome Y but also an effect of X on the mediator M. As such both Y and M can be written as linear functions of their predictors and normally distributed error terms,
The equations for Y and M that are used to define the indirect effect can be combined to obtain the total effect of X on Y under model 2.
From the structural model, note that ϵ_1 represents the error term for M and ϵ_2 represents the error term for Y. Assume that ϵ_2 is independent of M and of ϵ_1 and let

$\text{var}\,\epsilon_2 = \sigma_Y^2$ and $\text{var}\,\epsilon_1 = \sigma_1^2$

Using the specification of Y and M under Model 2 and under the assumption that Y and M have a multivariate normal distribution: Based on the multivariate normal distribution, the probability density function of Y and M can be written as

$$f(Y, M/X, \gamma_1, \gamma_2, \beta) = (2\pi)^{-n}(\sigma_Y^2\sigma_M^2)^{\frac{-n}{2}} exp\frac{-\nu_{11}}{2}\sum_{i=1}^{n} a_i^2 - \nu_{12}\sum_{i=1}^{n} a_i b_i - \frac{\nu_{22}}{2}\sum_{i=1}^{n} b_i^2 \quad (25)$$

Where a_i = y_i - $(\gamma_1 + \gamma_2\ \beta)x_i$ and $b_i = M_i$ - βX_i
The probability density function of Y and M can be viewed equivalently as the joint likelihood of γ_1, γ_2, and β.Therefore, from the likelihood we can obtain maximum likelihood estimates (MLEs) for γ_1, γ_2, and β. Once the MLEs have been calculated, they can be substituted into the likelihood to obtain estimated or profile likelihoods for the parameters of interest, namely γ_2 and β. To fnd the MLEs, we can maximize the log of the likelihood

$$L(\gamma_1, \gamma_2, \beta/X, M, Y, , \sigma_Y^2, \sigma_M^2) = -nln2\pi - \frac{n}{2}ln\sigma_Y^2, \sigma_M^2 - \frac{\upsilon_{11}}{2}\sum_{i=1}^{n} a_i^2 - \upsilon_{12}\sum_{i=1}^{n} a_i b_i - \frac{\upsilon_{22}}{2}\sum_{i=1}^{2} b_i^2 \quad (26)$$

Next, the derivative of the log likelihood with respect to the variable that is being maximized must be set equal to zero. Then, the value for the variable that solves the equation is considered the maximum likelihood estimate. We now obtain the MLE for γ_1

$\frac{\partial l}{\partial \gamma_1}$ = υ_{11} $\sum_{i=1}^{n}$ a_i x_i + υ_{12} $\sum_{i=1}^{n}$ b_i x_i

$= \frac{1}{2\sigma^2 Y}$ $\sum_{i=1}^{n}$ [y - $(\gamma_1 + \gamma_2\ \beta)x_i$] x_i - $\frac{\gamma_2}{\sigma^2 Y}$ $\sum_{i=1}^{n}$ $(M_i$ - $\beta x_i)$ x_i

$= \frac{1}{2\sigma^2 Y}$ $\sum_{i=1}^{n}$ y_i x_i - $\frac{\gamma_1}{2\sigma^2 Y}$ $\sum_{i=1}^{n}$ x_i^2 - $\frac{\gamma_2\beta}{2\sigma^2 Y}$ $\sum_{i=1}^{n}$ x_i^2 - $\frac{\gamma_2\beta}{2\sigma^2 Y}$ $\sum_{i=1}^{n}$ $m_i\, x_i$ + $\frac{\gamma_2\beta}{2\sigma^2 Y}$ $\sum_{i=1}^{n}$ x_i^2

$= \frac{1}{\sigma^2 Y}$ $\sum_{i=1}^{n}$ y_i x_i - $\frac{\gamma_1}{\sigma^2 Y}$ $\sum_{i=1}^{n}$ x_i^2 - $\frac{\gamma_2}{2\sigma^2 Y}$ $\sum_{i=1}^{n}$ m_i x_i

$= \frac{1}{\sigma^2 Y}$ $\sum_{i=1}^{n}$ $(y_i$ - γ_2 $m_i)\, x_i$ - $= \frac{\gamma_1}{\sigma^2 Y}$ $\sum_{i=1}^{n}$ $x_i^2 = 0$

$$\gamma_1 \ \sum_{i=1}^{n} x_i \ = \ \sum_{i=1}^{n} \ (y_i \text{-} \ \gamma_2 \ \ m_i \,) \, x_i$$

$$\hat{\gamma_1} = \frac{\sum_{i=1}^{n}(y_i - \gamma_2 m_i)x_i}{\sum_{i=1}^{n} x_i^2} \tag{27}$$

Using a similar process, we can obtain the MLE for γ_2

$$\frac{\partial l}{\partial \gamma_2} \ = \ \upsilon_{11} \ \sum_{i=1}^{n} \ \beta \, a_i \, x_i \ + \ \upsilon_{11} \ \sum_{i=1}^{n} \ a_i \, b_i \ \text{-} \ \upsilon_{12} \ \sum_{i=1}^{n} \ \beta \, b_i \, x_i \ \text{-} \ \upsilon_{12} \ \sum_{i=1}^{n} \ b_i^2$$

$$= \ \frac{\beta}{\sigma^2 Y} \ \sum_{i=1}^{n} (a_i \ \text{-} \ \gamma_2 \ \ b_i \,) \, x_i \ + \ \frac{1}{\sigma^2 Y} \ \sum_{i=1}^{n} (a_i \ \text{-} \ \gamma_2 \ \ b_i \,) \, b_i$$

$$\frac{1}{\sigma^2 Y} \ \sum_{i=1}^{n} (a_i \ \text{-} \ \gamma_2 \, b_i \,)(\, \beta \, x_i \ \text{-} \ b_i \,)$$

Replacing a_i and b_i in the equation above, we have

$$\frac{1}{\sigma_Y^2} \sum_{i=1}^{n} y_i \ \text{-} \ \gamma_1 \, x_i \ \text{-} \ \gamma_2 \, m_i \ = 0$$

Recall that

$$\hat{\gamma_1} = \frac{\sum_{i=1}^{n}(y_i - \gamma_2 m_i)x_i}{\sum_{i=1}^{n} x_i^2} \tag{28}$$

We can substitute in the MLE for γ_1 into the equation to obtain γ_2, the MLE for γ_2, which will not depend on γ_1

$$\sum_{i=1}^{n} \ y_i \, m_i \ \text{-} \ \sum_{i=1}^{n} \ (y_i \ \text{-} \ \gamma_2 \ \ m_i \,) \ x_i \ \sum_{i=1}^{n} \ x_i \ \sum_{i=1}^{n} \ \text{-} \ \hat{\gamma_2} \ \sum_{i=1}^{n} \ m_i^2 \ = 0$$

$$\sum_{i=1}^{n} \ y_i \, m_i \ \text{-} \ \frac{\sum_{i=1}^{n} y_i x_i \sum_{i=1}^{n} x_i m_i}{\sum_{i=1}^{n} x_i^2} \ + \ \hat{\gamma_2} \ \ \frac{\sum_{i=1}^{n} x_i m_i}{\sum_{i=1}^{n} x_i^2} \ \text{-} \ \hat{\gamma_2} \ \sum_{i=1}^{n} \ m_i^2 \ = 0$$

$$\sum_{i=1}^{n} \ y_i \ \ m_i \ \text{-} \ \frac{\sum_{i=1}^{n} y_i x_i \sum_{i=1}^{n} y_i m_i}{\sum_{i=1}^{n} x_i^2} \ \ \hat{\gamma_2} \ (\ \frac{(\sum_{i=1}^{n} x_i m_i)^2}{\sum_{i=1}^{n} x_i^2} \ \text{-} \ \sum_{i=1}^{n} \ m_i^2) = 0$$

$$\frac{\sum_{i=1}^{n} y_i m_i - \frac{\sum_{i=1}^{n} y_i x_i \sum_{i=1}^{n} x_i m_i}{\sum_{i=1}^{n} x_i^2}}{\sum_{i=1}^{n} m_i^2 - \frac{(\sum_{i=1}^{n} x_i m_i)^2}{\sum_{i=1}^{n} x_i^2}} \tag{29}$$

which simplifies to

$$\hat{\gamma_2} = \frac{\sum_{i=1}^{n} y_i m_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} y_i x_i \sum_{i=1}^{n} x_i m_i}{\sum_{i=1}^{n} m_i^2 \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i m_i)^2} \tag{30}$$

4 Results

4.1 The general Linear regression analysis

Before carrying out Linear regression analysis, we did the correlation test that is used to determine whether there is a relationship between dependent variable (HIV outcome)

and independent variables (education values) in the model. For instance the test to be carried out is

(i) H_0 : $\rho = 0$ (There is no correlation (r) between two variables)

(ii) H_1 : $\rho \neq 0$ (There is correlation (r) between two variables)

The correlation test results between independent variables and dependent variable are shown in Table 1.

Table 1: Correlation between dependent and independent variables

Model Estimate results	**estimate**	**S.E**	**Est/S.E**	**r-values**	**P-VALUE**
HIV Outcome	0.625	0.054	11.656	0.8461	0.0000
Education level	0.581	0.059	9.855	0.8893	0.0000

In Table 2, the results show that the independent variable is correlated with the dependent variable. This is indicated by p-values that are less than α, where α value is 0.05. Therefore there exist a relationship between independent (education level) variables with dependent variable (HIV outcome).
Next we carried out analysis to test the goodness of fit of the measurement model so as to assess the extent to which the latent variable were represented by the indicators (abstinence, use of condoms and non-marital sexual behaviour). The results are presented in Table 2

Table 2: Overall fit of model parameters

variable	**estimate**
Chi-square	21.695
p-value	.00002

A Chi-Square goodness of fit test is used to determine whether or not a categorical variable follows a hypothesized distribution. From Table 2, the results show a p-value < 0.05 which means that the data provide sufficient evidence against the null hypothesis, so we conclude that the observed data follow a different distribution than the theoretical one.
To compare two or more models i.e, measurement model and theoretical model, we used the Akaike Information Criterion(AIC) where AIC with smaller value of standardized estimates represents a better fit of the hypothesized model. The results are presented in Table 3.

Table 3: Goodness of fit

variable	**AIC**
Measurement model	1.178
Theoretical model	2.695

From Table 3, the results show that the measurement model has a better fit compared to theoretical model.

4.2 Measurement Model

The measurement model included one latent factor (Sexual Behaviour Patterns) and three observed variables(Use of Condoms, Abstinence and Non Marital Sexual behaviours). This model was estimated using the data 2018 KAIS which was a complete data since we used listwise deletion for missing data. Univariate skewness values ranged from –.07 to –1.12 and univariate kurtosis values ranged from –.79 to .91 confirming that the variables were indeed normally distributed as shown by West [22] and that no special estimators to address non normality were necessary. The overall fit indices and parameter estimates, taken directly from the output, are shown in Table 4 and 5. The tables present results in the form of regression equations, where the first set of equations is the unstandardized parameter estimates, and the second set is the standardized parameter estimates.

Table 4: overall fit and significance of model parameters

RMSEA	**estimate**
Chi-square	21.695 (2df)
Probability value for the chi-square statistic	.00002
Chi-square for this ML solution	20.686.
Comparative fit index (CFI)	.907
standardized RMSR	.044
Estimate	0.068
90 percent CI	0.345 0.419
Probability RMSEA	0.05 0.000

Table 5: Standardized parameter estimates

Sexual Behavior Patterns BY	**estimate**	**S.E**	**Est/S.E**	**P-VALUE**
Use Of Condoms	1.000	0.000	999.0	0.000
Abstinence	0.933	0.039	23.74	0.000
Non Marital Sexual Behaviors	0.880	0.038	23.057	0.000

From the results in Table 5, the overall significance of model parameters shows that the estimates of the model parameters were well represented by its indicators. An initial test of the measurement model revealed a very satisfactory fit to the data: χ^2 (21.685, N = 5000) = 113.22, p <.001; RMSEA = .055; SRMR = .044; and CFI = .907. All the factor loadings for the indicators on the latent variables were significant (p < .001), indicating that the latent factor was well represented by its indicators. From the results in Table 5 abstinence had a higher estimate 0.933 as compared to Non Marital Sexual behaviours with an estimate of 0.880 when use of condoms was controlled.

4.3 Structural Model

In Table 6 and 7, output of model parameters is shown. Table 6 provides four columns of output for each estimated parameter. The first column depicts the unstandardized coefficient; the second column, the standard error; the third column, the z-score demonstrating the significance of each parameter and the fourth column the standardized solution. These results can also be depicted in the output diagram in Figure 3

Table 6: Parameter estimates of structural parameters

INTERCEPTS	**estimate**	**S.E**	**Est/S.E**	**STD**
HIV Outcome	0.301	0.067	4.473.0	0.142
USE 0f Condoms	0.385	0.074	5.219	0.221
Abstinence	0.075	0.087	0.861	0.389
Non Marital Sexual Behaviors	0.126	0.084	1.495	0.135

From Table 6, Abstinence has a higher significance 0.861 as compared to other indicators, while non marital sexual behavior has a lower starndardised coefficient. The results further show that the disturbance terms (errors in prediction) were correlated among each construct, it shows there was a significant indirect effect (mediated) of the sexual behavior pattern on HIV outcome. Looking at their estimates the indirect effect is statistically significant but the direct is not. The reliability of the mean from the results can be deduced from the Standard Error results which show that small values of SE indicates that the sample mean is more accurate and a reflection of the actual population mean.

Table 7: Fit Indices for structural model

RMSEA	**estimate**
Estimate	0.071
90 percent CI	0.345 0.419
CFI	0.96
TLI	0.710
Probability RMSEA	0.05 0.000

In Table 7 the model shows appropriate fit according to all indices. They provide the estimates of the latent trait factor with a mean of (94.342) and variance of (101.86) as well as the estimate of the measurement error variances under residual variances values ranging between 9.389 to 11. 958.

5 Conclusion

In this study, we presented methods for developing the model and the fitting of the model to KAIS data in order to determine the appropriate probability distribution for data set with the aim of establishing the strength of the association between the categorical covariate(s) and the binary response through latent continuous mediator variable(s) using the structural equation modeling. The direct, indirect and total effects were examined in

a single model. Specifically, the sexual behavior patterns (intermediate variable) was used to explain why an independent variable (education level) affects the outcome HIV(status). From the results we find that there is a relationship between education attainment and knowledge /awareness of HIV/AIDS. In most research findings a relationship between high level of education and awareness of HIV has been observed such that those who have attained higher levels of education happen to be just as knowledgeable in HIV in terms of transmission, prevention , infection and control: However our analysis suggests that education levels is not associated with HIV prevalence after controlling for a number of socio-demographic characteristics and behavioral factors, unlike with other diseases higher levels of education do not appear to be protective against HIV positivity.

References

[1] Kharsany B.M and Karim A.Q (2016)HIV infection and AIDS in Sub-Saharan Africa: Current status , challenges and opportunities. *Open AIDS J* 2016;10:34-48.

[2] joint United Nations Programmes on Hiv/AIDS. (2014). The Gap Report. *ISBN 978-92-9253-064-4*

[3] National AIDS and STI Control Programme. (2018). Kenya HIV Estimate Report October 2018

[4] Abdool Karim S.S, Abdool Karim Q, Baxter C. (2015) Antibodies for HIV prevention in young women. *Curr Opin HIV AIDS.*2015 May;10(3):183-9. doi: 10.1097.

[5] Hayes A. F., Preacher K. J. and Myers T., A. (2011). *Mediation and the estimation of indirect effects in political communication research.* In E., P. Bucy and R.L.Holbert (Eds), Sourcebook for Political Communication Research: Methods, measures and analytical techniques pp (443-465) New York: Routledge.

[6] Muthen Bengt (1984). A A general structural equation model with dichotomous, ordered and continous latent variable indicators. *Psychometrica volume 49 Issue 1* pp 115-132.

[7] Winship C. and Mare R., D, (1983). Structural Equations and Path Analysis for Discrete Data. *american journal of sociology(internet)*

[8] Anna L. Hotton, Robert Garofalo, Lisa M. Kuhns, and Amy K. Johnson. (2013). Substance use as a mediator of the relationship between life stress and sexual risk among young transgender women. *AIDS Education and Prevention, 25(1), 62–71*

[9] Dohrenwend B.P (2000). The role of adversity and stress in Psychopathology. *J Health Soc Behav*

[10] Bryan et al. (1996) Increasing condom use: evaluation of a theory-based intervention to prevent sexually transmitted diseases in young women. *Healthy Psychology*

[11] Aminger and Kusters (1988) *Latent trait models with indicators of mixed measurement level* Springer science and business media.

[12] Fritz M.S.,et al (2016). The Combined Effects of Measurement Error and Omitting Confounders in the Single Mediator Model. *Multivariate Behav Res*, 51(5: 681-697.

[13] Loehlin John. C (1992) *Latent variable models: An introduction to factor, path and structural analysis 2nd ed.*Database: PsycINFO.

[14] Tabachnick B.G., and Fidell S.L.(2007) *Using Multivariate Statistics 2nd ed.* California State University

[15] Bentler,P.M (1980). Multivariate Analysis With Latent Variables.*Annual Review of Psychology*31(1):419-456

[16] Mackinnon,P.David (2008).*Introduction to Statistical Mediational Analysis.* Taylor and Francis goup,LLC.

[17] Mackinnon D.P.,Krull J.L and Lockwood C.M (2000).Equivalence of the mediation,confounding and suppression Effect. *Prevention Science*, 1(4): 173-181.

[18] Judd M.C.,and Kenny D.A. (1981). PROCESS ANALYSIS:Estimating Mediation in Treatment Evaluatios: it Evaluation Review, vol.5 No.5.

[19] Angela B., Sarah J.S., Michelle R.B 2006Mediation Analysis in HIV/AIDS Research:Estimating Multivariate Path Analytic Models in a Structural Equation Modeling Framework.*AIDS and Behaviour* 11, 365-383.

[20] Baron, R.M and David, A.K (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and Statistical considerations. *Journal of Personality and Social Psychology.* 51(6):1173-1182

[21] Kenny D.A (2018).Mediation. *power analysis app Medpower*

[22] West G Finch, J F and Curan P J. (1995). *Structural Equation Model with non Normal variables:Problems and Remedies*Publications, Inc.