A Logistic Regression Model to Identify Factors Influencing Secondary School Students' University Attendance Decision in the Southern Part of Sierra Leone

Regina Baby Sesay

[1]*Department of Mathematics and Statistics, School of Technology, Njala University, Njala, Sierra Leone*

***Abstract*-University education does not only increase earnings by providing skills that increase the chance of employment opportunities, but also promotes the economic growth of the country concern. For a developing country like Sierra Leone, a step to attain university education is a step towards moving away from poverty. It is therefore the desire of each and every parent to see their children through university level education. However, there are some unavoidable factors that may influence the ability and desire of provincial secondary (high) school students to further their education to university level. In Sierra Leone today, despite the increasing number of school dropouts, no clear-cut research has been carried out to address this great issue of concern. This research work, therefore, used a binary logistic regression modelling technique to identify the main factors influencing students' decision to attend a university of their choice after secondary (or high) school education. For this purpose, a stratified random sampling method was employed to select 363 respondents proportionately from each of the secondary schools in the Mokonde community, Korie chiefdom, Moyamba District, southern part of Sierra Leone. Data were collected from the selected respondents using structured questionnaires. To further ascertain the appropriateness of the chosen binary logistic regression model, two additional regression models, the proportional odd ordinal logistic regression model and the unconstrained partial proportional odd ordinal logistic regression model were also used in the analysis. However, statistical tests showed that the chosen binary logistic regression model outperformed the two ordinal regression models. Based on the result of the empirical analysis, the gender of the secondary (high) school student; the father's income level; the mother's income level; the annual average score and the number of study hours are the main factors influencing student's university attendance decision in the study area. Male secondary (or high) school students are more likely to attend university than their female counterparts. High school students whose fathers are on a high level income scale are more likely to attend university than those whose fathers are on the low level income scale. Also, the higher the average score of the secondary school student the greater the possibility of the student to enter university and above all, the more hours the student spends on studying his or her academic work, the greater the possibility for the student to enter university.**

***Keywords*: Logistic Regression; Predictors, Sierra Leone; Variance inflation Factor; Binomial Distribution, University Attendance**

I. INTRODUCTION

University is a place where people prepare for life as it impacts knowledge, and inject career values that contribute to long-term knowledge that may leads to a stable economic growth. In a developing country like Sierra Leone, where most of the inhabitants living in the provinces are subsistence farmers, university education can be the only means of escape from the poverty that is already evident in the lives of most of the provincial inhabitants. University education does not only increase earnings by providing skills that increase the chance of employment opportunities, it also promotes the economic growth of the country concern. In general, education and poverty are inversely related. This implies that, the higher the level of education (eg., university level) attained the lower the possibility of poverty in the life of the individual. Therefore, acquiring university education is one of the main steps towards achieving financial stability and a technologically improve economic growth. It is therefore, the desire of each and every secondary (or high) school student to proceed to the university after their secondary (or high) school education. However, there are some unavoidable factors influencing secondary (high) school students' university attendance decision. These include both academic and socioeconomic factors.

The socio-economic status of the family has a great impact on students' decision to further their education to university level. The social and economic status of students is generally determined by combining parents' qualifications, occupations and income standards [9]. Students from families with good financial and educational backgrounds tend to perform much better than those from poor and uneducated family backgrounds. Parents of students from high income and educational backgrounds can provide the latest technological facilities to enhance the educational capability of their children. Children whose families are of high educational scales have a statistically far better chance of participating in Tertiary Education [13]. In his research work, [1] also observed that in modern society, family influence played a very important role in the academic lives of students. Reference [4] even noticed in their findings that parent's

income or social status positively affects the student's test score in the examination

Also, the number of student's sturdy hours may influence the student's university attendance decision. Many studies have revealed that the study time has a strong relationship with the academic performance of the student (e.g., [7],[11]).

Again, the student's previous academic output which is reflected in the student's s class grade point average plays a significant role in the student's decision to attend (or not to attend) university. Students' academic performance measured by the student grade point average is a significant determinant of the educational pathway of the student. A student is normally placed into an academic or vocational stream based on his or her grade point average. Students placed in to the academic stream have high possibility of attaining university education.

Increasing provincial student's university attendance is considered as a step towards liberating the provincial inhabitants from poverty. Yet the number of studies in the literature with regards to high school students' university attendance decision in Sierra Leone is not known. This may be due to the fact that no clear cut studies have ever been carried out in Sierra Leone with regards to this topic.

According to [10], academic achievement affects the lives of the students to the full extent. This research work, therefore, used a logistic regression modelling technique to identify -the main determinants of students' university attendance decisions in the Mokonde, Njala community, Moyamba district, southern province of Sierra Leone as a case study,

Ц. MATERIALS AND METHODS

A. Theoretical Frameworks

This section focuses on the review of the theoretical and conceptual frameworks of using a logistic regression method for analyzing categorical outcome. It also points out the main statistics used in the logistics regression model checking.

1) Logistic Regression: In general, regression analysis is a predictive modeling technique used to investigates and estimate the relationship between a variable of interest called the outcome or dependent variable and one or more other variables that may influence the outcome variable. Regression analysis is used to model the relationship (or link) between an outcome variable and one or more predictor variable(s). There exist various procedures for fitting different types of regression models

Based on the type (or nature) of dependent variable, there exist different regression models used to establish relevant relation between a categorical dependent variable and one or more continuous or categorical independent variable(s). Table I presents a summary of the most commonly used regression models when the outcome variable is categorical.

Table I: Choice Of Regression Analysis For Categorical Outcome Variable

Outcome	Nature of Category	Variable Predictor	Regression
Categorical	Two Categories (Binary)	Continuous/ Categorical	Binomial Logistic Regression
Categorical	Multiclass categories with definite order	Continuous/ Categorical	Ordinal Logistic Regression
Categorical	Multiple categories with no definite order	Continuous/ Categorical	Multinomia l Logistic Regression

2) Binary Logistic Regression Model: Based on the number of independent variables and shape of the regression line, there exist different regression modelling techniques used to investigate relevant relationships and to make valuable predictions. Specifically, when the response variable is dichotomous, (i.e. has only 2 possible values), it is good to have a regression model that predicts the value as a probability score that ranges between 0 and 1. Therefore, based on the binary nature of the dependent variable (i.e. will attend or will not attend), this work used a binary logistic regression modeling technique to investigate the factors influencing high school student's decision to attend university after their high school graduation. The logistic regression is a transformation of the linear regression using the sigmoid function.

3) The logistic sigmoid function: Due to the dichotomous nature of the outcome variable in a logistic regression, unlike the linear regression (with an output consisting of continues number of values), the logistic regression transforms it output using logistic sigmoid function. The logistic sigmoid function fits a set of data with independent variable(s) taking any real value, and the dependent variable being either o or 1. Therefore, the logistic sigmoid function maps predicted values to the probabilities as it maps any real value into another value between 0 and 1 This result to a perfect output representation of the probabilities as the function always lies in the range of 0 to 1. For the parameter, say **z,** the logistic sigmoid function is given as:

$$\boldsymbol{h(Z) = \frac{1}{1 - e^{-z}}}$$

Where:

- h(z) = output between 0 and 1 (probability estimate)
- z = input to the function (your algorithm's prediction e.g. mx + b)
- e = base of natural log

Figure 1 presents an example of a plot of the sigmoid function

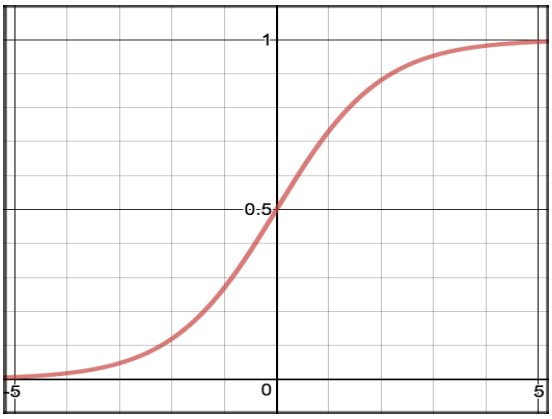

Figure I: The plot of the sigmoid function. with y=0.5 at x=0

The Binomial Distribution as an Error Distribution in logistic Regression

The binomial distribution is appropriate to use as an error distribution in logistic regression because: the outcome of interest is dichotomous (a success or a failure); and number of independent trials were considered in obtaining the required information for the analysis-. For the purpose of this research work, let

$$Y_i = \begin{cases} 1 \text{ if the ith secondary school student will enter University} \\ 0 \text{ 1 if the ith secondary school student will not enter University} \end{cases}$$

where yi is the category of the university attendance (or entrance) decision for student i. Here, yi is considered as a realization of a random variable Yi that can take the values one and zero with probabilities pi and $1-pi$ respectively. The distribution of Yi is called a Bernoulli distribution with parameter pi and can be written as

$$\boldsymbol{pr(Y_i = y_i)} = \boldsymbol{p_i^{y_i}](1-p_i)^{1-y_i}}$$

For $\boldsymbol{y_i = 0, 1}$. If $\boldsymbol{y_i = 1}$ we obtain $\boldsymbol{p_i}$, and if $\boldsymbol{y_i = 0}$ we obtain $\boldsymbol{1 - p_i}$

The Expected value and variance of $\boldsymbol{Y_i}$ are

$$\boldsymbol{E(Y_i) = \mu_i = p_i} \text{ and } \boldsymbol{var(Y_i) = \sigma_i^2 = p_i(1-p_i)}$$

The mean and variance depend on the underlying probability $\boldsymbol{p_i}$ and any factor that affects the probability will alter not just the mean but also the variance of the observations. This suggest a linear model that allows For k independent high school student $\boldsymbol{y_1, \ldots, y_k}$, (k=1,…,). The $\boldsymbol{i^{th}}$ high school student can be treated as a realization of a random variable Y_i. We assume that Y_i has a binomial distribution $\boldsymbol{Y_i \sim Bin(n_i, p_i)}$ with binomial denominator $\boldsymbol{n_i}$ and probability $\boldsymbol{p_i}$. For individual data, $\boldsymbol{n_i = 1}$ for all i.

In this research work, the binomial distribution was used as an error distribution in the logistic Regression analysis because:

- the outcome variable (student's university attendance decision) used in the analysis is dichotomous.
- a number of independent trials were considered in the analysis.

4) The Logistic Regression Model Used in the Analysis: The best statistical model to use when the dependent variable is dichotomous is the binary logistic regression model. This model uses the logistic link function to model a binary dependent variable. The model for the logistic regression analysis is given as:

$$\boldsymbol{\log \frac{\hat{p}}{(1-\hat{p})} = \beta_0 + \beta_1 x_1 +, \ldots, \beta_p x_k}$$

$$\boldsymbol{Y \sim Binomial(\hat{p})}$$

Where $\boldsymbol{\hat{p}}$ is the predicted probability that $\boldsymbol{Y = 1}$, given the values of $\boldsymbol{x_1, \ldots, x_k}$,

the $\boldsymbol{\beta_i}$s are the parameters to be estimated by the logistic model and

$\boldsymbol{\log \frac{\hat{p}}{(1-\hat{p})}}$ is the logit link function that takes a linear combination of the covariate values (which may take any value between ±∞) and convert those values to the scale of probability (between 0 and 1)

5) *Parameter Interpretation:* Unlike the simple linear model $\boldsymbol{Y = \beta_0 + \beta_1 x_1}$ indicating that if x increases by 1, Y increases by $\boldsymbol{.\beta_1}$, in the logistic regression model, it is $\boldsymbol{\log \frac{\hat{p}}{(1-\hat{p})}}$ which increases by $\boldsymbol{.\beta_1}$.

Let the predicted probability of the event of interest be $\boldsymbol{p_0}$ when $\boldsymbol{x = 0}$ and $\boldsymbol{x = 1}$, Then we have

$$\boldsymbol{\log \frac{\hat{p}}{(1-\hat{p})} = \beta_0}$$

$$\boldsymbol{\log \frac{\hat{p}_1}{1-\hat{p}_1} = \beta_0 + \beta_1}$$

$$\boldsymbol{\log \frac{\hat{p}_1}{1-\hat{p}_1} = \log \frac{\hat{p}_0}{1-\hat{p}_0} + \beta_1}$$

Taking Exponent on both sides of this equation we have:

$$\boldsymbol{e^{\log\left(\frac{p_1}{1-\hat{p}_1}\right)} = e^{\log\left(\frac{\hat{p}_0}{1-\hat{p}_0}\right)+\beta_1}}$$

This gives

$$\boldsymbol{\frac{p_1}{1-\hat{p}_1} = \frac{\hat{p}_0}{1-\hat{p}_0} \times e^{\beta_1}}$$

This means, when x increases by 1, the odds of a positive outcome increase by a factor

of $\boldsymbol{e_1^{\beta}}$. Therefore, $\boldsymbol{e_1^{\beta}}$ is called the odds ratio for a unit increase in x.

6) Assumptions attached to the Binary Logistic Regression: Although the logistic regression does not impose stringent assumptions like homoscedasticity; linearity between the dependent and independent variables or normally distributed error terms, there are certain assumptions that must be

satisfied for the result of the binary logistic regression model to be valid. These assumptions include:

- The dependent variable must be categorical (Dichotomous)
- No multicollinearity, meaning the independent variables should not be highly correlated with each other.
- There should be a linear relationship between the link function, $\left(\log\left(\frac{p}{1-p}\right)\right)$ and independent variables in the logit mode

7) Logistic Regression Model Checking:

The goodness of fit tests help to find out if the model at hand is correctly specified. Therefore, before trusting the result of the binary logistic regression model, it is good to check the model beyond all reasonable doubt to make sure that the model we have assumed is correctly specified.

8) Deviance Goodness-of-Fit Test: In logistic regression analysis, the maximum likelihood estimation is used to compute the logistic regression estimates. This iterative process finds the minimal discrepancy between the observed response, Y, and the predicted response, Yˆ . The resulting summary measure of this discrepancy is the -2 loglikelihood or -2LL, known as the deviance [12]. The bigger the deviance, the bigger the discrepancy between the observe, and expected values. Adding more predictors to the model will result to a smaller deviance, indicating an improvement in fit. The model deviance with one or more predictors is compared to a model deviance without any predictors, called the null model. The likelihood ratio test is used to compare the deviance of the null model (denoted as $\boldsymbol{L_0}$) and the deviance of the full model (denoted as $\boldsymbol{L_1}$)

$$\boldsymbol{G^2 = \text{deviance}L_0 - \text{deviance}L_1}$$

$$\boldsymbol{= -2ln\left(\frac{L_0}{L_1}\right) = [-2\ln(L_0)] - [-2\ln(L_1)]}$$

where:

$\boldsymbol{G^2}$ is distributed as a chi-squared value with df equal to the number of predictors added to the model.

9) Pearson Goodness-of-Fit Test:

The Pearson Goodness-of-Fit test is mainly and frequently intended to be used for categorical variables (or discrete) variable. This goodness-of-fit test compares the observed values to the expected (fitted or predicted) values.

The Pearson chi-square is calculated as:

$$\boldsymbol{\chi^2 = \sum_j \frac{(O_j - E_j)^2}{E_j}}$$

Where:

Oi = the number of observed items in category i,

Ei = number of expected items in category i. (In order to use this test, each value of Ei must be at least 5.)

If the fitted model is correct, this statistic has approximately a chi-square distribution,

Note:

- the expected number of items in a category is determined by the expected value of a binomial random variable. $E_i = np_i$ with n being the number of observations, and p_i, the probability of obtaining an observation in category i.
- In order to use this test, each value of E_i must be at least five (5).

The Pearson's Goodness-of-Fit Test is a right-tailed test. Therefore, a value of $\chi 2=0$, at the extreme left end of the distribution, would be equivalent to a perfect fit.

10) Measures of the Predictive Power of the Logistic Regression Model: In logistic regression, the R^2 is mostly use for two main purposes:

- to get a measure of how well the dependent variable can be predicted based on the independent variables.
- to test whether the model needs additional nonlinearities and interactions to correctly represent the data.

Different methods exist for calculating R^2 for logistic regression. Two of these methods, one proposed by McFadden (1974) and another by Cox and Snell (1989) are presented below

McFadden R^2

Let $\boldsymbol{L_0}$ be the value of the likelihood function for a model with no predictors, and let $\boldsymbol{L_M}$ be the likelihood for the model being estimated. The McFadden's R^2 is defined as

$$\boldsymbol{R^2_{McF} = 1 - \frac{\ln(L_M)}{\ln(L_O)}}$$

where ln(.) is the natural logarithm

It is important to note that in using, $\boldsymbol{R^2}$, adding ***any*** variable may tends to increase it value, even if that variable is irrelevant. For this reason, the adjusted $\boldsymbol{R^2}$ was used in this work to access the predictive power of the model.

Cox and Snell R^2

Similarly, for *the Cox and Snell R^2, let L_0 be the value of the maximized likelihood for a model with only the intercept (with no predictor) and let L_M be the likelihood of the estimated model with all the predictors. Then* the Cox and Snell R^2 is

$$\boldsymbol{R^2_{C\&S} = 1 - \left(\frac{L_O}{L_M}\right)^{\frac{2}{n}}}$$

where *n* is the sample size

III METHODOLOGY

This section introduces stages involved in the data analysis. It also points out the type of data analysis adopted at each stage together with the need for each analysis.

A. Study Area

The study was carried out in the Mokonde Community, Kori Chiefdom, Moyamba district in the Southern part of Sierra Leone. The Njala Mokonde Community is dominated by secondary school (High School) children most of whose parents are either traders or university workers.

B. Population and Sample size: The target population consisted of

all high school students in the sturdy area. The sample size was chosen based on the work of

[14].for sample size consideration in logistic regression analysis

C. Sampling and Data Collection

A random sampling technique was used to select three hundred and sixty-three (363) secondary (high) school students from the communities in the study area. Questionnaires containing questions relating to student's university attendance (or entrance) decisions together with potential factors that might influence the type of decision were administered to all the selected secondary (or high) school students-. The data obtained provided information on the personal, academic and socioeconomic factors influencing the decision of secondary school students to attend (or enter into) any university of their choice.

IV. EMPIRICAL ANALYSIS

A. Descriptive and Exploratory Data Analysis.

Table II: Dependent (Dv) And Independent (Iv) Variables To Be Modeled

Variable Name	IV/DV	Valid Range	Variable Type
Student's University Attendance Decision	**DV**	Will attend University Will not Attend University	Categorical (Dichotomous**)**
Father's Educational Level	IV	No Formal Education Primary Education Secondary Education Tertiary Education	Categorical **(**Ordinal**)**
Mother's Educational Level	IV	No Formal Education Primary Education Secondary Education Tertiary Education	Categorical **(**Ordinal**)**
Mother's Income Level	IV	Low Medium High	Categorical **(**Ordinal**)**
Father's Income Level	IV	Low Medium High	Categorical **(**Ordinal)
Gender	**IV**	Female Male	(Dichotomous**)**
Study Hours		**1-4 hours**	Continuous
Average Score	IV	**Percentage**	Continuous

1) Descriptive Statistics

Table III: Descriptive Statistics For Categorical Variable

	N	Range	Minimum	Maximum
Gender	363	1.0	0.0	1.0
Father's income level	363	2.00	1.00	3.00
Mother's income level	363	2.00	1.00	3.00
Valid N (listwise)	363			

Table Iv Descriptive Statistics For Continuous Variable

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
Number of Sturdy Hours	363	.70	4.00	2.5515	.88237	.779
Average score for accademic year	363	32.00	89.00	63.5014	16.14187	260.560
Valid N (listwise)	363					

2) Exploratory data Analysis (EDA)

Preliminary exploratory data analysis helped to identify useful relationships between the dependent variable and each of the independent variables. EDA provides useful information on the associations which, can be later incorporated into the Logistic regression model. For this reason, bivariate exploratory analysis was carried out to know if there was a relationship between the continuous independent variables and the categorical outcome variable. The independence sample t-test was used to explore the relationship between each of the continuous independent variables and the outcome variable, students university attendance decision. The common assumptions made when doing the t-test were considered. The assumption of the t-test for independent means focuses on sampling, research design, measurement, population distributions and population variance. The t-test for independent means is considered typically robust for violations of the normal distribution assumption (with a larger sample size). However, this work used the QQ-plot to see if the assumption of normality was satisfied before using the t-test.

- *Quantile-Quantile (Q-Q) plot for continuous independent variables*

The Q-Q plot is a graphical method for comparing two probability distributions by plotting their quantiles against each other. A concave departure from the straight line in the Q-Q plot is an indication of a heavy tailed distribution, whereas a convex departure is an indication of a thin tail.

Based on the Q-Q plots for the continuous variables presented in figures 2 and 3, it is clear that, the distributions of the continuous independent variables are not perfectly normally distributed. However, because of the central Limit Theorem (sample size is greater than 30) and the data was obtained randomly, the t-test was carried out.

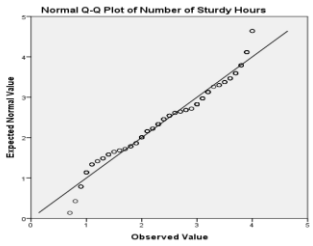

Figure 2: O-Q plot of number of study hours

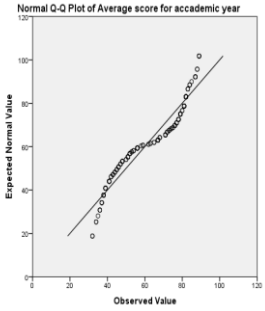

Figure 3: O-Q plot of Annual Average Score

The independent sample t-test was carried out for each of the continuous independent variables, to determine if:

1) there is a statistically significant difference in the mean sturdy hours of student who would like to attend university after high school and those who would not want to attend university after high school.

2) there is a statistically significant difference in the mean annual score for students who would like to attend university after high school and those who would not want to attend university after high school.

The independence sample t-tests helped to determine if the binary logistic regression model was a good choice model for the variables used in the analysis. A significant difference in mean, implies, running a logistic regression would be the best, as the results would be significant. Tables V and VI present the outputs of the independent sample t-tests for the continuous variables used in the binary logistic regression model

From Table V, the P-value for the independence sample t-test for the number of sturdy hours in relation to whether a student will attend a university or not is far below the threshold significance level of 0.05 for the two-tailed tests. This means that the mean difference in the number of study hours for those high school students who would want to attend university and those who would not want to attend university after high school graduation is statistically significant. This further implies that, there is a relationship between the number of study hours and high school student decision to attend (or not to attend) university. To further explore this relationship, a logistic regression model was used in the analysis.

Similarly, from Table VI, the significance level in the independence sample t-test, for number of sturdy hours is below the threshold significant level of 0.05. This means that the mean difference in the annual average score for those high school students who would like to attend university and those who would not like to attend university after high school graduation is statistically significant. This further implies that, there is a relationship between the annual average score and high school student decision to attend (or not to attend) university, To further confirm this relationship, a logistic regression model was used in the analysis

Table V: Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Number of Sturdy Hours	Equal variances assumed	6.099	.014	2.151	361	.032	.33163	.15418	.02843	.63482
	Equal variances not assumed			1.841	40.394	.043	.33163	.18014	-.03235	.69560

Table VI : Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Average score for accademic year	Equal variances assumed	3.313	.045	6.512	361	.000	17.48547	2.68513	12.2050	22.7659
	Equal variances not assumed			6.901	44.393	.000	17.48547	2.53383	12.3801	22.5908

A. The Binary Logistic Regression Analysis

The logistic regression output (in TableVII) helped to summarize the significance of the independent variables individually whilst controlling for the other independent variables in the model. In particular, the logistic regression output in TableVII helped to identify those independent variables that positively (or negatively) influence secondary (or high) school students' university attendance decisions. The output also determines how each determinant influenced students' university attendance decision.

From table VII, the father's income level with p-value (for Wald) that is less than the threshold significance level of 0.05 for all levels was one of the main factors influencing high school students' university attendance decisions. The father's income level was first tested as a whole (Father_Incom) and then 1st and 2nd income levels compared to the reference category 3, ($High\ income\ level$). The odd ratio (Exp (B)) for father's middle income level (Father_Incom(2)) is 5.798. This implies that those high school students whose fathers are in the middle-income level are 5.798 times more likely to attend university than those whose fathers are in the low income level.

For gender, the odd ratio compares the likelihood of a male high school student's university attendance (entrance) to that of their female counterparts. The odds (Exp(B)) for 'male high school students' university attendance are 3.086 times higher than for 'female high school students. The odds for 'female high school students' university attendance is 1/3.086 = 0.324. This implies that females were 0.324 times less likely to attend university than their male counterparts.

The mother's income level is another significant factor that determines student's university attendance decision.

The wald's P-value for mother's income level (Mother_Incom) as a whole is less than the chosen significant level of 0.05.

Also, the p-value for Mother_Incom(1) (i.e. mother's income level low) is less than the chosen significance level of 0.05

and the odd ratio (Exp(B)) associated with Mother_Incom(1) is 4.436. This implies that 'high school students whose mothers are on the low income level are 4.436 times more likely to attend university compared to the reference category

(those whose mothers are on the high income level). In other words, the odds of increase in university attendance for high school students increase for those high school students whose mothers are on the low income scale.

Again, the independent variable, number of study hours (Stu_Hrs) was found to be a very significant factor in determining students' university attendance decisions. Its parameter estimate is positive with a significant level of 0.005. Its odd ratio is greater than one. This implies that, the odds of an increase in students' university enrollment are higher for students who spend more time studying (more study hours) than those who spend less time studying. That is, the probability of an increase in students' university attendance increases with a unit (hour) increase in number of study hours than at original.

Finally, the average score of the student was another significant factor in determining student's university attendance decision. Its parameter estimate is positive with a significant level of $0.001 < P - value$. It odd ratio is greater than one (i.e $5.228 > 1$). This implies that, the odds of an increase in students' university entrance or attendance decision are higher for students with high average score than those with low average score. That is, the probability of students' university attendance increases with a unit (percentage) increase in the average score of the student.

Table VII Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Gender(1)	1.127	.318	12.575	1	.000	3.086	1.655	5.752
	Father_Incom			31.171	2	.000			
	Father_Incom(1)	1.384	.497	7.773	1	.005	3.992	1.508	10.564
	Father_Incom(2)	1.758	.321	29.962	1	.000	5.798	3.090	10.879
	Mother_Incom			10.791	2	.005			
	Mother_Incom(1)	1.490	.458	10.558	1	.001	4.436	1.806	10.894
	Mother_Incom(2)	.570	.330	2.979	1	.084	1.769	.926	3.380
	Stu_Hrs	.629	.181	12.034	1	.001	1.876	1.315	2.677
	Ave_Score	1.654	.478	11.981	1	.001	5.228	2.049	13.339
	Constant	.275	.705	.152	1	.696	1.317		

a. Variable(s) entered on step 1: Gender, Father_Incom, Mother_Incom, Stu_Hrs, Ave_Score.

1) Test of logistic regression model Coefficients: The hypothesis tested for the significance of the logistic regression model coefficients were stated as:

$H_0: \beta_i = 0 \qquad H_A: \beta_i \neq 0 \qquad for\ i = 1,2,3,4,5$

The model coefficients contained in the column headed B of Table VII can be positive or negative.

A negative value means that the odds of university attendance decreases. However, for this binary logistic regression model, the column headed B from table consist of only positive values that are each significantly different from zero (p-values are each < 0.05) except for one category of mother's income level, Mother_Incom(2) with P-value = .084>0.05 . However the moder's income a whole is a significant predictor of the dependent variable, 'high school student's university attendance decision'.

Therefore, the null hypothesis for the test of the binary logistic model coefficient (i.e., $H_0: \beta_i = 0$) was rejected in favor of the alternative hypothesis ($H_A: \beta_i \neq 0$).

The full model being tested is:

$$\ln\left[\frac{\hat{P}(x)}{1-\hat{P}(x)}\right] =$$

$$0.275+1.127GenderMale+1.384\text{Father's\_Income}low+1.758\text{Father's\_Income}Middle+1.490\text{Mother's\_Income}low+0..570\text{Mother's\_Income}Middle+0..629StuHrs+1.654(Ave\_Score)$$

B. Model Checking

1) *Chi-square goodness of fit test for model coefficients*: The omnibus test presented in table VIII was used to check if the present (new) model with explanatory variables included is an improvement over the baseline model. A significantly reduced value of the Log- likelihoods (-2LLs) suggests that the new model is explaining more of the variation in the outcome variable than the baseline (or constant only) model. From Table VIII, the chi-square statistic is highly significant (chi- square= 99.378 df=8, p<.000). This shows that, the present (new) model is significantly better compared to the baseline model.

The null and alternative hypotheses for the overall Chi-square test were stated as:

$\boldsymbol{H_0: \beta_i = 0 \qquad for\ all\ i}$

$\boldsymbol{H_A: \beta_i \neq 0}$ **for at least one coefficient**

The null hypothesis, H_0 was rejected since p-value = .000. is less than the chosen significant value of 0.05.

Table VIII Omnibus Tests Of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	99.378	8	.000
	Block	99.378	8	.000
	Model	99.378	8	.000

2) Model Validation (Classification table): The Classification table from output result summarizes the observed group and the predicted group classification. From Table IX, it can be seen that the present logistic regression model correctly classified the outcome for about 90% of the cases.

TABLE IX: CLASSIFICATION TABLE

Observed			Predicted		
			University Enrolment		Percentage Correct
			will not attend University	will attend University	
Step 1	University Enrolment	will not attend University	2	34	5.6
		will attend University	0	327	100
Overall Percentage					91

3) Model chi-square goodness of fit test: The hypothesis tested for the model goodness of fit was stated as:

*H*0: The model is a good fitting model.

Ha: The model is not a good fitting model.

From Table **X**, the tests of goodness of fit shows that, the model is a good fit to the data as

$p = 0.348 > 0.05.$

Table X: Hosmer And Lemeshow Test

Step	Chi-square	df	Sig.
1	8.931	8	.348

4) Measures of the Predictive Power of the Model: The model summary result presented in Table **XI:** shows that, between 33% and 49% of the variability in students' university attendance (entrance) decisions can be explained by the present logistic regression model.

Table XI: Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	143.540[a]	.326	.485

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

5) Influential Observations and Outliers: One of the main assumptions of logistic regression is that there are no extreme outliers or influential observations in the dataset. It is, therefore good to find out if there are observations that do not fit the model well (outliers), have strange values (leverage) or that have undue influence on the model (influential). For this purpose, this work used the cook's distance, denoted as Di to identify points that negatively affect the logistic regression model. The measurement is a combination of each observation's leverage and residual values. The higher the leverage and residuals, the higher the Cook' distance. An observation (a data point) that has a large Di value indicates that it strongly influences the fitted values. More specifically, a Di value greater than 1 indicates that an influential observation is present.

The maximum and minimum values of the Cook's Distance for the present analysis are presented in Table **XII**. From Table **XII** , the maximum value of Di is 0.17550 which is less than one (<1). Therefore, the issue of influential observation or outlier is not a concern in this analysis.

TABLE XII: DESCRIPTIVE STATISTICS OF THE COOK'S INFLUENCE STATISTICS

	N	Range	Minimum	Maximum	Variance
Analog of Cook's influence statistics	363	.17546	.00003	.17550	.0259712
Valid N (listwise)	363				

6) The ROC curve: The ROC curve was created by plotting the sensitivity (or true positive rate) against the specificity (true negative) at various threshold settings. First, we calculate sensitivity and specificity pairs for each possible cutoff point and plot sensitivity on the y axis by 1-specificity on the x axis. The resulting curve is called the receiver operating characteristic (ROC) curve.. The ROC curve was not only use to measure the predictive accuracy of the present logistic regression model but to also evaluates how well the present logistic regression model classifies positive and negative outcomes at all possible cutoffs. The area under the ROC curve ranges from 0.5 and 1.0 with larger values indicative of better fit. The diagonal line in the curve represents chance. The curve presented in figure 4 is well above the diagonal line. This was further justified by the output presented in Table XIII, which contains the area under the ROC curve together with the inference statistics about the curve. From Table XIII, the area under the curve (AUC) is 0.847, with 95% confidence interval (.785, .910). The area (under the curve) is significantly different from 0.5 as the $-value = 0.000 < 0.05$. This is a justification that the logistic regression classified the group significantly better than by chance. This further represents a high predictive accuracy of the chosen model. In other words, an AUC value of 0.847 (which is close to 1) indicates that the model reliably distinguished between ' high school 'student who will enter University' and those who 'will not enter university'.

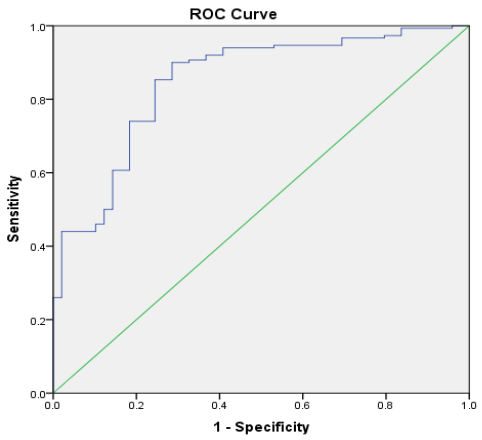

Figure 4: Receiver Operating Characteristic (ROC) curve

Table XIII: Area Under the Curve

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.847	.032	.000	.785	.910

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

7) Multicolinearity: Multicolinearity causes the precision of the coefficient estimates to reduced, which makes the p-values

very unreliable. Hence making it difficult to decide on which predictor variables are statistically significant. This work, therefore, used the value of the tolerance and its reciprocal, called the variance inflation factor (VIF) to detect multicollinearity in the data set. The variable's tolerance is $1 - R^2$. If the value of tolerance is less than 0.2 or 0.1 and, simultaneously, the value of the VIF is 10 and above, then multicollinearity is problematic.

For the present logistic regression analysis, the highest value of, R^2 presented in Table XVII: is the Negelkerke, R^2 with a value equal to 0.485. The tolerance is calculated as $1 - R^2 = 1 - 0..485 = 0.515$and it VIF is 2.0681 $(i.e., \frac{1}{R^2} = \frac{1}{0.485} = 2.0618.)$ The tolerance is far above 0.1 and the value of VIF is far below 10. It was therefore concluded that multicolinearity is not problematic. In addition, the standard errors of the coefficients are not too significant. This further suggested that multicolinearity is not an issue in the present data set.

C. The Ordinal Logistic Regression Analysis

For comparative purpose, the proportional odd ordinal logistic and the unrestricted partial proportional odd ordinal logistic regression analysis were also carried out on the same data set with different reordering of the dependent variable. Unlike the binary logistic regression model where the dependent variable, students' university attendance decision was binary (as 'attend' and 'not attend'), for the ordinal logistic regression analysis, the dependent variable, high school student university attendance decision was ordered as: 'more likely to attend', 'likely to attend' and 'not likely to attend'

1) Proportional odd ordinal logistic regression analysis: Table XIV Presents the regression output for the proportional odd ordinal logistic regression model. From the output presented, all the significant independent variables in the chosen binary logistic regression model were also significant in the proportional odd ordinal logistic regression model. However, the brant test presented in Table XV showed that the proportional odd ordinal logistic regression model is not a good fitting model for the data set, as the major assumption, the Parallel regression assumption was violated for the whole model.

Table XIV: Proportional Odds Ordinal Logistic Regression

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept):1	-11.9752	1.3093	-9.146	< 2e-16 ***
(Intercept):2	-18.7327	1.7824	-10.510	< 2e-16 ***
Gender	0.7995	0.3394	2.356	0.0185 *
Father_Incom	1.5235	0.2866	5.315	1.07e-07 ***
Mother_Incom	0.6416	0.2553	2.513	0.0120 *
Study_Hrs	1.7463	0.2318	7.534	4.91e-14 ***
Ave_score	2.5622	0.2656	9.646	< 2e-16 ***

- *Test for Proportional odds*

This research work used the brant test to test whether the observed deviations from the fitted ordinal logistic regression model are larger than what could be attributed to chance alone, The **r**esult of the brant test presented in Table XV was used to test if the parallel assumption (proportional odd assumption) of the proportional odd ordinal logistic regression holds in the present model. The proportional odd assumption is that, the relationship between each pair of outcome groups is the same. The null hypothesis for the brant test is that the parallel assumption holds. This means that the parallel assumption only holds for variables with p-values greater than the chosen significant value of 0.05. The brant test tests f**or** both the individual variables and for the whole model. The omnibus variable in the brant test output stands for the whole model. From the output presented in TableXV the parallel (or proportional odd) assumption does holds for the whole model as the p-value for the omnibus variable is less than the chosen significant value of 0.05. Also, the P-value for two of the independent variable, Gender and Ave_score are each less than the chosen significant value of 0.05. This rendered the proportional odd ordinal logistic regression model unfit to be used in the analysis. For this purpose, the unconstrained partial proportional odds ordinal logistic regression model, which does not wholly depend on the proportional odd assumption was used as an alternative model in the ordinal logistic regression analysis.

Table XV: Brant Test For The Parallel Assumption Of The Proportional Odd Model

Test for	X2	df	probability
Omnibus	21.24	5	0
Gender	6.78	1	0.01
Father_Incom	0.04	1	0.84
Mother_Incom	0.31	1	0.58
Study_Hrs	1.23	1	0.27
Ave_score	7.73	1	0.01
H0: Parallel Regression Assumption holds			
Test for	X2	Df	Probability
Omnibus	21.24434801	5	0.0007283113
Gender	6.77999599	1	0.0092185097
Father_Incom	0.04325689	1	0.8352423181
Mother_Incom	0.30791734	1	0.5789610076
Study_Hrs	1.22986478	1	0.2674333369
Ave_score	7.72885925	1	0.0054345074

Table XVI presents regression estimates together with their corresponding standard errors and P-values for both the binary logistic regression and the unrestricted partial proportional odds ordinal logistic regression models. This table allows for easy comparison to be made between the two models in terms

of coefficients, p-values and standard errors. Most of the significant independent variables for the proposed binary logistic regression model are also significant for the unrestricted partial proportional odds ordinal logistic regression model. However, the standard error for some of the independent variables (e.g. Gender, Father_Incom(2), Stu_Hrs) are considerably lower for the binary logistic regression model as compared to the unrestricted partial proportional odd ordinal logistic regression model.

Table XVI: Comparing Outputs For The Binary Logistic Regression And The Unconstrained Partial Proportional Odds Ordinal Logistic Regression Models

Independent Variables	Model coefficient (B)		Standard Error (SE)		P-value	
	Binary Logistic Model	partial proportional odd OR Model	Binary Logistic Model	partial proportional odd OR Model	Binary Logistic Model	partial proportional odd OR Model
Gender(Male)	i.127	2.0306	0.318	0.6460	0.00	0.001669 **
Father_Incom(1)	1.384	1.5241	0.497	0.4516	0.005	0.000739 ***
Father_Incom(2)	1.758	1.3747	0.321	0.4115	0.00	0.000836 ***
Mother_Incom(1)	1.490	0.7923	0.458	0.4332	0.001	0.008544 **
Mother_Incom(2)	0.570	0.9759	0.330	0.3711	0.084	0.067436
Stu_Hrs	0.629	1.6147	0.181	0.3450	0.001	2.86e-06 ***
Ave_Score	1.654	2.2657	0.478	0.3436	0.001	4.26e-11 ***

- Model comparison for the Binary Logistic, proportional odds ordinal logistic and Unconstrained partial proportional odds ordinal logistic regression Models:

Table XVII presents the Akaike information criteria (AIC) and the residual deviance used to compare the binary logistic regression, proportional odds ordinal logistic regression and the unconstrained partial proportional odd ordinal logistic regression models:

The values of the AIC and residual deviance for the three models showed that the proposed binary logistic regression model with the least AIC and residual deviance values performed better than the proportional odds and unconstrained partial proportional odds ordinal logistic regression models

Table XVII: AKAIKE INFORMATION CRITERIA (AIC) AND RESIDUAL DEVIANCE

Model	AIC	Residual deviance
Binary Logistic Regression	135.33	123.33
Proportional odds ordinal logistic regression	250.408	250.408
Unconstrained partial proportional odds ordinal logistic regression	249.149	225.149

V. RESULTS AND DISCUSSION

This research used a logistic regression analysis to identify the main factors influencing the decision of secondary (or high) school students to enter university after their high school education.

The study was carried out in the Njala Mokonde community located in the Korie Chiefdom, Moyamba district, Southern part of Sierra Leone. The main aim (or purpose) of the study was to identify the main factors influencing secondary (or high) school students' university attendance decision and to determine the effect of each factor on the high school students' decision to attend (or enter) University.

To achieve this aim, a binary logistic regression modelling technique was used in the empirical analysis.

The dependent variable, 'high school students' university attendance decision', was dichotomous as it was classified as 'will attend (or enter) university' and 'will not attend university'. The model for the binary logistic regression was significant, as the test of the full model against a model with only the intercepts was statistically significant. This showed that the predictors as a set reliably distinguished between students in their various decisions to attend university after high school or not. (chi square = 99.378, $p < .05$ with df=8). The pseudo R^2 values (e.g. Negelkerke.=0 .485=49%) presented in Table **XI:** indicates that the logistic regression model with its independent variables explained about 50% of the total variability in students' decisions to attend (or enter) university. To further prove the uniqueness of the chosen binary logistic regression model, two other regression models, the proportional odd and the unrestricted partial proportional odds ordinal logistic regression models were also used in the analysis. However, statistical tests showed that the present binary logistic regression model outperformed both the proportional odds and the unrestricted partial proportional odds ordinal logistic regression models

-Several factors were initially considered as potential determinants of students' university attendance decision. However, the result of the analysis showed that, the gender, the father's income level (Fa_Incom); the mother's income (Mo_Incom) level; the average score and the number of study hours (Stu_hrs) were the main factors influencing students' university attendance decision.

The father's income level was first tested as a whole (Father_Incom) and then 1st and 2nd income levels compared to the reference category 3, $(High\ income\ level)$. The odd ratio (Exp(B)) for Father's middle income level (Father_Incom(2)) was 5.798. This implied that those high school students whose fathers are in the middle-income level were 5.798 times more likely to attend University than those whose fathers were in the low income level. This is in line with other research findings that, low-income families with parents who have little or no education may create a less enthusiastic atmosphere concerning education and children's futures ([2], [6]). Reference [8] also reported that, one in six Japanese children lives in poverty and the prospect of acquiring a good education is often hampered by their parent's inability to finance their schooling to High School.

For gender, the odd ratio compares the likelihood of a male high school student's university attendance (enrollment) to that of their female counterparts. The odds (Exp(B)) for 'male high school students' are 3.086 times higher than for 'female high school students. In other words, the odds for 'female high school students' university attendance is 1/3.086 = 0.324. This implies that females were 0.324 times less likely to attend university than their male counterparts. Reference [3] indicated in their findings that safety concerns can prevent children from attending school, particularly girls. This research finding is also in line with the view that, girls continue to fare much worse than boys in terms of school reention, completion, and performance [15].

The mother's income level is another significant factor that determines student's university attendance decision. The wald's P-value for mother's income level (Mother_Incom) as a whole is less than the chosen significant level of 0.05. Also, the p-value for Mother_Incom(1) (i.e. mother's income level low) is less than the chosen significance level of 0.05 and the odd ratio, (Exp(B)) associated with Mother_Incom(1) is 4.436. This implies that 'high school students whose mothers are on the low income level are 4.436 times more likely to attend university compared to the reference category (those whose mothers are on the high income level). In other words, the odds of increase in university attendance for high school students increase for those high school students whose mothers are on the low income scale. Reference [4] also discovered in their research findings that parent's income or social status positively affects the students' tests scores in examinations.

Again, the independent variable, number of study hours (Stu_Hrs) was found to be a very significant factor in determining students' university attendance decisions. Its parameter estimate is positive with a significant level of 0.005, Its odd ratio is greater than one. This implies that, the odds of an increase in students' university attendance are higher for students who spend more time studying (more study hours) than those who spend less time studying. That is, the probability of an increase in students' university attendance increases with a unit (hour) increase in number of study hours than at original. Reference [5] even observed that, students who are very successful in their desired career have longer study time.

Finally, the average score of the student was another significant factor in determining student's university attendance decision. Its parameter estimate is positive with a significant level of $0.001 < P - value$. Its odd ratio is greater than one (i.e $5.228 > 1$). This implies that, the odds of an increase in students' university entrance or attendance are higher for students with high average scores than those with low average scores. That is, the probability of student's university attendance increases with a unit (percentage) increase in the average score of the student.

VI. CONCLUSION

A Logistic regression analysis was carried out to identify the main factors influencing secondary (or high) school students' university attendance decision after high school education. The statistical tests (e.g., the chi-square goodness of fit test) showed that the logistic regression model was a good fit for modelling the data. To further justify the appropriateness of the binary logistic regression model, two additional regression models, the proportional odds ordinal logistic regression model and the unrestricted partial proportional odds ordinal logistic regression models were also used in the analysis. Statistical tests (AIC and residual deviance) showed that the binary logistic regression model fitted much better than the two ordinal logistic regression models. Several factors were initially considered as potential determinants of high school student's university attendance decision. However, the result of the logistic regression analysis showed that, the gender, the father's income level (Fa_Incom); the mother's income (Mo_Incom) level; the annual average score and the number of study hours (Stu_hrs), were the main factors influencing secondary school students' university attendance decision.

VII. RECOMMENDATIONS

1) Among the significant determinants of student's university attendance decision identified from the binary logistic regression analysis, the number of sturdy hours is the only factor used in the analysis that a student may have absolute control over. That is, a student can decide to attain university level education by simply increasing the number of sturdy hours. As increase in the number of study hours may result to an excellent university entrance grade and will therefore, increase the likelihood of proceeding to a choice university after the secondary (high) school education. Therefore, based on the research findings, the researcher, strongly recommend

that students should increase the number of study hours in order to increase their chance of University admission after high school.

2) Future Research Direction: The present research work only focused on secondary school pupils living in the study area. It is therefore, recommended that further research be carried out to include the upper primary school pupils in the sturdy. Also, similar study is recommended to be carried out in the capital city of Sierra Leone, where majority of the petty traders hocking on the street, especially the female youths are school dropout who for one reason or the other could not have access to University education.

REFERENCES

[1] Ahawo, H. (2009). Factors Enhancing Student Academic Performance in Public Mixed Day

[2] Bell, K. L, Allen, J. P., Mauser, S. T., & O'Conner, T. G. (1996). Family factors and young adults transitions: Educational attainment and occupational prestige..

[3] Colclough C, Rose P and Tembon M (2000) Gender inequalities in primary schooling: The roles of poverty and adverse cultural practice. *International Journal of Educational Development* 20(1): 5–27

[4] Considine, G. &Zappala, G. (2002). Influence of social and economic disadvantage in the academic performance of school students in Australia. Journal of Sociology, 38, 129-148.

[5] D. E. Ukpong & I. N. George (2013), Length of Study-Time Behaviour and Academic Achievement of Social Studies Education Students in the University of Uyo, International Education Studies; Vol. 6, No. 3; 2013 ISSN 1913-9020 E-ISSN 1913-9039

[6] Galambos, N. L., & Silbereisen, R. K. (1987). Income change, parental life outlook, and adolescent expectations for job success. Journal of Marriage and the Family, 49, 141-149.

[7] Gbore, A. (2006). *Factor Wondering Effective Study Habits Among Students: A Hand Book for Students in Colleges and Universities*. Nakuru: Egerton Publishing

[8] Hagiwara Y, Reynolds I. (2015) In Japan, 1 in 6 children lives in poverty, putting education, future at stake. The Japan Times. September 10, 2015;Sect. National/Social Issues. images/0019/001907/190771.

[9] Jeynes, W. H. (2002). Examining the effects of parental absence on the academic achievement of adolescents: The challenge of controlling for family income. *Journal of family and Economic Issues*, *23* (2), 65-78.

[10] Koç, Y., Terzioğlu, E. A., Kayalar, F. (2018). Examination of individual achievement motivation and general self-efficacy of candidates entering into special talent exam in physical education and sports sciences. Journal of Sports and Performance Researches, 9(2), 64–73.

[11] Logunmak in, G. F. (2001). Predicting the academic success of students from diverse populations. Journal of College Student Retention, 2(4), 295-311.

[12] McCullagh P, Nelder JA (1989), Generalized Linear Models New York: Chapman and Hall

[13] Oloo, M.A. (2003). Gender disparity in student Achievement in day secondary schools. Migori: Maseno University

[14] Peduzzi, P, Concato, J, Kemper, E, et al. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996; 49: 1373–1379

[15] Sabates R, Akyeampong K, Westbrook J, et al. (2011) *School dropout: Patterns, causes, changes and policies*. Education for All Global Monitoring Report. Available at: http://unesdoc.unesco.org/