THE UNIVERSITY OF NAIROBI (UoN)

ANALYSIS OF KCSE PERFORMANCE IN NAKURU COUNTY: A GENERALIZED ESTIMATING EQUATIONS APPROACH.

By

ELVIS KARANJA MUCHENE

I56/67573/2013

Supervisor:

DR. NELSON O. OWUOR

This Project is submitted in partial fulfillment of the requirements for the degree of Master of Science in Social Statistics in the School of Mathematics, Chiromo Campus.

June, 2015

DECLARATION

This Project is my original work and has not been presented for a degree in any other University.

Signature ..... Date.....

ELVIS KARANJA MUCHENE

This Project has been submitted for examination with my approval as the university supervisor.

Signature ..... Date.....

DR. NELSON OWUOR

Dedication

To my loving and caring family, friends, colleagues at work, lecturers and campus classmates who were all very instrumental in the success of this project.

Acknowledgements

I am sincerely grateful to God for his blessings and for successfully seeing me through this academic journey. I would like to specifically thank my supervisor Dr. Nelson O. Owuor for his guidance throughout this project. Its through his support and advice that this work turned out a masterpiece. Additionally I wish to acknowledge the dedicated support and guidance from my brother Leacky Kamau, who endlessly and tirelessly offered academic guidance in writing this project. He played a key role in making this project a success especially in learning statistical softwares necessary for this study. I am deeply humbled and very grateful to my fiance Grace Wamaitha for her patience, love, support and encouragement she offered throughout my academic journal. Finally I will forever be grateful to the generous support and encouragement from my family throughout my studies.

Elvis Karanja Muchene,

June 16, 2015

Nairobi, Kenya

Abstract

In the Kenyan education system, progression in tertiary education is dependent on a standardized national examination administered by the Kenya National Examinations Council (KNEC). The ministry of education guidelines stipulates that the pass mark for the university entry examination is C plus and above. A student who scores C+ or higher is eligible for direct admittance to university program. Publicly available data on Kenya Certificate of Secondary Education (KCSE) performance in Kenya for the years 2006-2010 was analyzed. Differences between the different school types (boys only, girls only, or mixed schools) as well as differences in performance between boys and girls were assessed. A generalized estimating equations marginal model was applied in order to account for association between scores within a school in the five year period using the SAS procedure PROC GENMOD. Flexibility in the trend was captured by additional quadratic and cubic time effects. GEE goodness of fit statistics, the quasi-likelihood under independence model criterion (QIC) was used to select best mean model as well as best working correlation structure for the study. Finally contrasts of interest were performed. A model with school, gender specific intercepts and common slopes was selected with exchangeable correlation structure. Results indicated that there was a significant difference between the different school types in their candidates probability of attaining the stipulated minimum university entry grade. In particular, boys only schools had the highest probability, followed by girls only schools and finally mixed schools. Moreover contrasts indicated that boys in boys only schools had a higher success rate than boys in mixed schools. Girls in girls only schools had a higher success rate than girls in mixed schools while boys in mixed schools performed better than girls in mixed schools. The success rate in KCSE however did not depend on the year under review as was evident in the linear, quadratic and cubic slope parameters which were not statistically significant.

Key words: *Exchangeable correlation,Generalized estimating equations, KCSE, QIC,*

Contents

List of Tables

List of Figures

List of Abbreviations

AIC	Akaike Information Criterion
GEE	Generalized Estimating Equations
GLM	Generalized Linear Model
KCSE	Kenya Certificate of Secondary Education
KNEC	Kenya National Examination Council
MLE	Maximum Likelihood Estimation
QIC	Quasi-Likelihood under the Independence model Criterion

Chapter 1

INTRODUCTION

In the Kenyan education system, progression to tertiary education is dependent on a standardized examination administered by the Kenya National Examination Council (KNEC). The examination administered leads to the award of the Kenya Certificate of Secondary Education (KCSE). The ministry of education guidelines stipulates that the pass mark for KCSE is a mean grade of C plus (commonly denoted C+) and above, which corresponds to a minimum of six points on a twelve point grading scale, with the twelve points corresponding to the highest possible score. A student who scores C+ or higher is deemed eligible for direct admittance to a university program.

1.1 Study Background

The performance in the KCSE examination varies across the country depending on many factors including; the classification of the schools as either national, county, the number of candidates in a school, whether the school is boys only, girls only or of mixed gender school, available facilities for teaching, location of school in terms of political stability in the region amongst a myriad of other factors.

Until the year 2014, the release of KCSE results-which is usually done every year around February-March through a ceremony headed by the minister of education- included ranking of the students' performance individually, (best 100 candidates in each province and nationally by gender) as well as the ranking of schools based on the mean grade of the schools' candidates. This ranking mostly stimulated healthy competition amongst schools in a bid to outperform each other in the subsequent examinations. Some schools were consistent over the years in terms of their ranking while 'one time wonders' were also a common occurrence. However, there has not been much reported analysis or comparison of schools performance taking into account the potential effect of time. Moreover at face value, the ranking popularly reported by the ministry of education does not form a good scientific basis for comparison of performance across boys only, girls only or mixed schools.

1.2 Problem Statement.

The Kenyan government is committed to provision of equal access to secondary education to all Kenyans. In line with this, the government launched free day secondary education in 2008 as was stipulated in the Kenya Education Sector Support Programme (KESSP) which was launched in 2005. This led to 1.7 Million students benefitting from the programme in the year,(Njoroge and Kerei, 2012). Several studies have been undertaken in establishing difference in performance between males and females in the Kenya Certificate of Secondary Education (KCSE) with no specific inference to the type of schools the students belong especially in Nakuru County. There is need to further extend such studies to establish how secondary school performance in KCSE in Nakuru county varies from different school types namely Boys only secondary schools, Girls only secondary schools and Mixed secondary schools (Boys & Girls). This study will focus on establishing if there exists difference in secondary school performance between the three school types, as well as a study on difference in performance across similar genders between different school types, in Nakuru County.

1.3 Study Objectives

1.3.1 General Objective:

Gain insightful analysis on KCSE performances in Nakuru County while taking into account different school types and gender over time, sufficient enough to warrant need for interventions from Nakuru county government, ministry of education as well as other relevant stakeholders.

1.3.2 Specific Objectives:

1. Establish if there exists a significant difference in overall KCSE performance between Mixed schools, Boys Schools & Girls schools in Nakuru County.
2. Establish if boys performance differs significantly between mixed schools and boys schools, enough to warrant for interventions from relevant education bodies.
3. Establish if girls performance differs significantly between mixed schools and girls schools, enough to warrant for interventions from relevant education stakeholders.
4. Establish if there exists a significant difference in overall KCSE performance between boys & girls in mixed school.

1.4 Significance of the study.

This study will provide insights on KCSE performances in Nakuru County. It will focus on the relationship between gender of the students as well as the type of school the students belong to relative to their performance in KCSE over time. The insights gained from this study can be adopted by relevant education stakeholders in Nakuru County as well as the national government

in formulating policies geared towards an improved performance in the national examinations within the county.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

Student performance in secondary education can vary widely due to several factors. The approach adopted for this paper is generalized estimating equations (GEE) of Zeger et al. (1988) which is an extension of generalized linear models (GLM) to longitudinal data analysis and uses quasi-likelihood estimation. With their approach, we do not only look at annual KCSE performance independently, but try to account for potential correlation in the KCSE performance indicators within one school over time.

2.2 Standard GEE theory

Various approaches exist for analyzing longitudinal data sets with the mixed effects models which use full likelihood approaches and generalized estimating equations GEEs which use quasi-likelihood approach as the most commonly used. GEE models population averaged profiles whereas the mixed effect models both the fixed effects, i.e. population averaged profiles as well as the random effects (subject/individual) specific profiles. According to Molenberghs and Verbeke (2005), whereas full likelihood models have the benefit of gaining efficiency, this

comes at an extra cost of an increased rate of model misspecification due to the computational complexity they entail. A full likelihood procedure can however be replaced by quasi-likelihood especially if one is interested in the first order marginal mean parameters as well as the pairwise associations. This leads to the generalized estimating equations, usually denoted GEE as proposed by Zeger et al. (1988).

Generalized estimating equation are an analysis method and not necessary a model. Instead, as noted by Weaver (2009) , a model to be fitted using GEE approach is specified through a link function that relates mean response to a regression equation while assuming distributional assumptions for the response. A working correlation is also specified.

Liu (2010) Showed that the first extension of GEE, usually denoted GEE1, requires only the correct specification of the univariate marginal distributions and in estimating the main effect parameters, the information of the association structure is not used. This yields consistent main effect estimators even when the association structure is misspecified. Molenberghs and Verbeke (2005) however noted that, severe misspecification of the association structure may affect the efficiency of GEE1 estimators, and as a result are less adequate if the study interest is largely based on the association of parameters. On the other hand the second order extension of GEEs, denoted GEE2, includes both the marginal pairwise associations as well as the correlations. GEE2 is as efficient as the full likelihood approach, but as Molenberghs and Verbeke (2005) observed, bias is likely to occur in estimation of the main effect parameters when the association structure is misspecified. In our study, we focus on GEE1.

The choice of association structure depends on the type of study being conducted since GEE uses several correlation structures to model the correlation matrix among observations within each cluster. Compound symmetry/exchangeable correlation structure assumes equal correlations within subjects of interest at all the time points in the model. Under the assumption of independence of observations within a cluster, the correlations are assumed to be zero which

is the usual assumption in classical logistic regression. On the other hand, the first order auto regressive correlation structure denoted AR (1) assumes that adjacent observations have higher correlations than the no adjacent ones. Finally, the unstructured assumes different correlations amongst observations, and is thus estimated independently from the data, (Lawal, 2003).

Lawal (2003) introduced two different time concepts in employing GEE method. He defined a time stationery covariate as a between subject covariate that would be repeated in each of the time point measures, for every subject. In our study, this time stationery covariate is the variable gender which is repeated every year for every school. The time varying covariate on the other hand, is a within subject variate and assumes different values for each of the time point measure on each subject. In our study, this can be viewed as the KCSE grade which varies each year for each school.

GEE has the property of yielding consistent main effect estimators even when the association structure is misspecified implying that the point estimates as well as the standard errors are asymptotically correct. These standard errors are popularly known as the robust standard errors and the variance estimator is referred to as the sandwich estimator. Standard errors in GEE are reported in the form of the empirically corrected standard errors and the model based standard errors. As noted by Molenberghs and Verbeke (2005), its of no use to report on the model based standard errors since they are generally incorrect, unless they would be of scientific interest on the study where they can be looked at as an indication of the distance between the working assumptions for the correlation and the true structure. The empirically corrected standard errors are however of interest to the study in GEE analysis. Whereas a far apart distance between both standard errors may be an indication of poor choice of working assumptions, we recall that a poor working assumption is not wrong. Lawal (2003) further noted that a robust estimator is deemed a good estimate if the number of clusters is large. Naive variance estimates are on the other hand correct if the correlation has been correctly modelled.

Longitudinal datasets are prone to missing observations especially due to incidences of drop outs amongst other factors. Just like in any other analysis, the issue of missing data and how to handle it is of key interest and deserves special treatment even in the context of GEE models. Data is considered to be missing completely at random (MCAR) when the probability of missingness is completely independent of the outcome, whether missing or observed. On the other hand, data is considered missing at random (MAR) when the probability of missingness is independent of the vector of missing outcomes, but may be dependent on observed outcomes.For GEE estimator, valid inferences can be obtained from data which is MCAR or MAR,(Zorn, 2001).

In conclusion, some of the benefits obtained from GEE as observed by Weaver (2009) and Ghisletta and Spini (2004) include but not limited to the fact that it accounts for within subject correlations, allows for time varying covariates as well as irregularly timed measurements, allows for a range of correlations and can be applied to incomplete data as long as the individual observations are missing completely at random. GEEs have no strict distribution assumptions and instead assume the variance of the outcome variable to be expressed as a function of the expectation. The GEE approach can also be easily implemented in several statistical softwares especially in SAS using the PROC GENMOD procedure. On the other hand, GEE also has its limitation just like any other statistical approach. As observed by Ghisletta and Spini (2004), the technique is asymptotic and requires large sample sizes for unbiased and consistent estimation.

2.3 Review of GEE applications.

GEE has widely been used in the context of Gaussian and non-Gaussian correlated datasets. Ghisletta and Spini (2004) noted that this approach has widely been applied in biological, pharmacological and closely related disciplines with its application in educational and social sciences still being quite scarce despite longitudinal data in education and social sciences being a common

occurrence.

Molenberghs and Verbeke (2005) presented several case studies where application of GEE in the medical practice and epidemiological studies was illustrated. He used clustered data from the developmental toxicological area conducted under the U.S National Toxicology Programme (NTP) to fit a model in standard GEE approach using the SAS procedure GENMOD. The study aim was to model the effect of a five doses/chemicals in mice. In this analysis, the working assumptions of independence and exchangeable were considered since the other assumptions such as AR (1) and unstructured were less sensible given the nature of the data. The analysis compared model based and empirically corrected standard errors and there was a clear difference in the case of independence working assumptions and much less in the case of exchangeable assumption. Molenberghs and Verbeke (2005) noted that as long as the study interest was to assess the effect of a dose, GEE1 would suffice but if there was additional interest in association, then GEE1 opt to be cautiously applied. Finally, the working assumption of exchangeability was deemed reasonable both on biological grounds and also putting into consideration the design of the study.

Lawal (2003) applied GEE approach on a six cities longitudinal study whose interest was to assess the health effects of pollution. In particular, the study centered on whether age has an effect on childs wheezing status. He analyzed data from two of the cities namely the Kingston-Harriman, which is considered a more polluted city, and Portage. Children between 7 years and 10 years were examined for wheezing/panting while also recording the mothers smoking habits at the start of the study. In this study, the response at each age was a childs wheezing status which was of binary nature of either zero if no wheeze and one if there was wheeze. Covariates of interest were the city, childs city of residence, which was also binary with 1=if child lived in Kingston-Harriman, and 0=if child lived in Portage. Other covariates of interest were the smoke status measured as the maternal cigarettes smoking at that age in packs per day, and the childs age in years. Age effect was assessed in linear, quadratic and cubic terms. The model was implemented in SAS software using the PROC GENMOD procedure.

Results from the study showed that the parameter estimates were similar for all the correlation structures. Childs city of residence was significant with the more polluted city (Kingston-Harriman) having a higher tendency to increase odds of wheezing by a factor of 1.65. Smoking was also significant with the odds of a child wheezing increasing by a factor of 1.68 for every two packs per day smoked by the mother. Contrasts were further performed to test for no age effect in the model. Results indicated lack of significance in the contrasts using any of the correlation models implying that there was no age effect.

Using data from the Swiss Interdisciplinary Longitudinal study on the Oldest Old (SWILSO-O), Ghisletta and Spini (2004) applied GEE approach in assessing predictors of drop out in a longitudinal study of an old Swiss sample assessed five times. SWILSO-O is a multi-cohort longitudinal study on psychological, health, social and sociological situation. Age, sex, living context, living arrangement, depressive symptoms, physical troubles and social economic status (SES) were the variables considered in the analysis. Results from the GEE showed that the time varying covariate age was statistically significant. The unstructured and non-stationery GEE models showed that the physical status was significant and thus meaningful to the participation of the study. SES as well as the living context also had a strong effect. From the results obtained after fitting different correlation structures, the study recommended the Logit GEE model with unstructured specification. This was informed by the fact that the unstructured model implied the drop out process between first and subsequent waves were different from successive drop out processes. Further, the model confirmed a known drop out background in ageing studies that older ages, lower SES as well as lower physical health status are strong predictors of drop out in such researches.

Within the education sector however, while the academic performance of students has been evaluated mostly via standardized examinations such as the British General certificate of Secondary Education (GCSE) and in the Kenyan context, the Kenya Certificate of Secondary Education

(KCSE), literature on retrospective studies on the possible trends in performance over time is limited. For instance, McManus et al. (2013) tackled the problem of continuity in performance of students in the medical school based on their secondary school performance. In particular, they compared data from five longitudinal studies of UK students and doctors between 1970 and early 2000s. Their meta- analysis however used correlation analysis and path diagrams.

Charnley (2008) in his report on accessing GCSE performance of independent pupils based on gender and school type differences, used data from the MidYIS (Middle Years Information System) project which is operated by the Curriculum, Evaluation and Management (CEM) center at Durham University to perform separate contrasts in performance between boys in boys school and in mixed schools as well as girls in girls schools and in mixed schools. Contrasts were performed on the basis of logarithmic regression equations which were separately computed for both genders. He observed that pupils in single sex schools performed significantly better in most of the subjects than pupils in mixed schools.

Eisenkopf *et. al* conducted an experiment in Switzerland to study the effects of random assignment to co-educational and single sex classes on academic performance of female high school students Eisenkopf et al. (2011). Estimation results showed that single sex improves performance of female school students in mathematics and this effect is more positive if the single sex class is taught by a male teacher.

In Kenya, Mburu (2013) conducted a study to investigate the influence of the type of school attended on gender differences in KCSE performance in Kericho and Kipkelion districts. The main objectives of the study was to establish if the social classroom interactions had an effect on male and female student academic performance, as well as the type of school attended. Two questionnaires were administered to teachers and students to collect data while descriptive statistics were used in data analysis. In particular Chi-square and correlations statistics were used in analyzing data on the influence of the type of school attended on students academic

performance and compared to the students results in KCSE.

Results from the study showed that the type of school attended was a determinant on gender difference in academic performance. Girls from girls schools only had a better academic performance compared to girls from mixed schools. The same case applied to boys where boys from single sex schools had a better academic performance than those from a mixed school. The study also revealed that majority of the students had a strong preference of joining single sex schools over the mixed schools. One of the factors attributed to poor performance of both genders from mixed schools compared to single sex schools was the number of distractions from opposite genders especially for the females.

Challenges faced by girl child in academic performance have been of interest to most stakeholders especially within the education sector as well as non-governmental organizations. For instance, Makewa et al. (2014) employed descriptive-comparative, correlation and cross section survey approach to study if there was any relationship in girl child challenges and academic achievement in mixed secondary schools in Mbooni district. In particular the study adopted a descriptive research design to identify the challenges faced by the girl child in mixed schools and correlation predictions were made on the effects of these challenges on academic achievement. Results from the study showed that there was a moderate negative correlation between girls performance and female teachers as role models.

Yara and Catherine (2011) used multiple linear regressions to assess the determinants of performance in mathematics during the KCSE within Nyamaiya division. A more advanced approach to analyzing secondary school education data was presented by Bagaka's (2011). Using multilevel data with data hierarchy at school- class-student level, he used a two-level hierarchical linear mixed-effects model. In addition, his analysis included the students KCSE performance as a predictor amongst other covariates of interest. The outcome of interest was the standardized questionnaire outcomes for teachers and students.

Chapter 3

METHODOLOGY

This chapter describes the data and variables that will be used to examine the objectives of the study as well as any data clean up or manipulation techniques that will be necessary. We will further describe the application of generalized estimating equations to this study given the data.

3.1 Data

Longitudinal data on Kenya Certificate of Secondary Education (KCSE) performance was obtained from the Kenyan governments open data website for the period 2006-2010. [1]. Longitudinal data consists of repeated measures/observations of an outcome variable for each experimental unit/subject, recorded over a period of time. For the purpose of this analysis, a unit/subject refers to a school within Nakuru district, Nakuru County for which we have results for KCSE for at least one year within the 5-year period under consideration. Each subject may have a set of covariates associated with them. One of the characteristics of the outcomes in longitudinal data is that outcomes from the same subject are usually correlated.

[1] `https://www.opendata.go.ke/Education/KCSE-Exam-Results-2006-to-2010/ycfy-7tnf`

We then apply appropriate filters to extract data for Nakuru district in Nakuru County. The success or failure in the exams is based on the grade attained in KCSE in each year.

3.2 Computing response variable.

The interest is to compute a binary response variable of either success or failure for each school within a given period of time based on the grade attained. We define an indicator variable such that, its a Pass if the KCSE mean grade is higher than or equal to C+ (C Plus) and Fail when the a KCSE mean grade is less than C+. The choice of this categorical outcome is informed by the fact that a minimum of C+ is the official Kenyan government's pass mark to join a university.

We thus have a binary indicator for school i in year j for gender k defined as below;

$$I_{ijk} = \begin{cases} 1, & if \geq C+ \\ 0, & if < C+ \end{cases} \tag{3.1}$$

In order to obtain the response of interest, the data is aggregated based on the indicator variable by calculating the total number of students of a particular gender in that school who passed or failed in each of the years. The final response variable is thus binomially distributed as follows;

$$Y_{ijk} \sim Binomial\ (n_{ijk}, \pi_{ijk}) \tag{3.2}$$

Where, Y_{ijk} is the number of students from school i who passed (had C+ and above) in year j for each gender k.

n_{ijk} is the total number of candidates of gender k in school i and year j obtained as the sum of number candidates who passed and those that failed in that particular year.

π_{ijk} is the probability of passing for a candidate of gender k in school i and year j.

This binomial response is measured repeatedly for each school. The available data however has gaps in some years whereby performance of some schools is not reported. The missingness pattern is assumed to be missing completely at random (MCAR) and therefore the analysis does not try to accommodate it.

Some of the assumptions on this variable are;

1. Y_{ijk} are not necessarily normal. Infact, they follow a binomial distribution.
2. Y_{ijk} are not necessarily independent. Measurements from the same school are correlated.

Let N be the number of subjects and n_i be the number of repeated measurements of the i^{th} subject. We can therefore group the response for the i^{th} subject into an $n_i \times 1$ vector as below;

$$y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ . \\ . \\ y_{in_i} \end{pmatrix} ;\ i = 1, 2, \ldots\ldots N \tag{3.3}$$

3.3 Computing covariates.

The covariates of interest in this study are gender, year and school type where school type is either single sex, i.e. Boys only or Girls only, or mixed school which comprises of both genders. However, the data obtained did not explicitly categorize schools into the three school types of

interest for this study. Instead the data only provides the performance for each gender in every school in each year.

Using this information, we define a dummy variable for the covariates for each gender in school i in year j as below;

$$
\begin{aligned}
X_{ij1} &= \begin{cases} 1, \text{if boys from boys only school} \\ 0, \text{otherwise} \end{cases} \\
X_{ij2} &= \begin{cases} 1, \text{if girls from girls only school} \\ 0, \text{otherwise} \end{cases} \\
X_{ij3} &= \begin{cases} 1, \text{if boys from mixed school} \\ 0, \text{otherwise} \end{cases} \\
X_{ij4} &= \begin{cases} 1, \text{if girls from mixed school} \\ 0, \text{otherwise} \end{cases}
\end{aligned}
\tag{3.4}
$$

We can similarly group the vector of covariates into an $n_i \times p$ matrix of covariates as below;

$$
X_i = \begin{pmatrix} x'_{i1} \\ x'_{i2} \\ \vdots \\ x'_{in_i} \end{pmatrix} = \begin{pmatrix} x_{i11} & x_{i12} & \cdots & x_{i1p} \\ x_{i21} & x_{i22} & \cdots & x_{i2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{in_i1} & x_{in_i2} & \cdots & x_{in_ip} \end{pmatrix} \tag{3.5}
$$

3.4 Marginal Model: Generalized Estimating Equations (GEE).

Generalized estimating equations, usually denoted GEE, are basically an extension of generalized linear models (GLMs) to accommodate correlation in outcomes. One of the properties of longitudinal data is that the outcomes of a single subject are usually correlated. GEE is a Population-Averaged models, usually denoted (PA), where the aggregate response for the population is modeled rather than modeling a subject specific profile like in the generalized linear mixed-effects models (random-effects models).

In this study, we apply the methodology for generalized estimating equations, (GEE) in order to account for the correlation between outcomes of the same school. We adopt GEE1 where one does not use information of the association structure to estimate the main effects parameter. GEE1 only requires the correct specification of the univariate marginal distribution.

As such, GEE1 yields consistent main-effect estimators, regardless of whether the association structure is mis-specified or not.

3.4.1 Model Specification.

One of the model assumptions in fitting GEE is that the covariates can be nonlinear transformations of the original independent variables, and can also have interaction terms. In this study, we perform transformations on the variable Year by centering it (subtracting 2006 from each year) so as to ease model convergence and to ensure that the model intercepts are meaningful. In this case, the model intercept corresponds to probability of success in the year 2006. Moreover, transformation of the centred year variable to account for quadratic and cubic effect on the outcome probability is performed. We also introduce an interaction term of gender with the school type (boys only, girls only or mixed school) to allow for contrasts between performances of similar genders in different school types.

Based on this information, we define a mean structure that comprises of intercepts specific for the school type, gender, interaction term as well as linear, quadratic and cubic time effects. We also incorporate the school specific slopes or common slopes and assess their appropriateness.

$$Y_{ijk} = \beta_{ok} + \beta_{1k} Year_{ij} + \beta_{2k} Year^2_{ij} + \beta_{2k} Year^3_{ij} \tag{3.6}$$

We therefore formulate a general model whose response Y_i is associated with a $p \times 1$ vector of covariates X_{ij} as below;

$$Y_{ij} = \begin{cases} \beta_{01} + \beta_{11}\, Year_{ij} + \beta_{21}\, Year^2{}_{ij} + \beta_{31}\, Year^3{}_{ij}\ , \ \ if\ Boys\ from\ boys\ Only\ schools \\ \beta_{02} + \beta_{12}\, Year_{ij} + \beta_{22}\, Year^2{}_{ij} + \beta_{32}\, Year^3{}_{ij}\ , \ \ if\ Girls\ \ from\ girls\ Only\ schools \\ \beta_{03} + \beta_{13}\, Year_{ij} + \beta_{23}\, Year^2{}_{ij} + \beta_{33}\, Year^3{}_{ij}\ , \ \ if\ Boys\ from\ Mixed\ Schools \\ \beta_{04} + \beta_{14}\, Year_{ij} + \beta_{24}\, Year^2{}_{ij} + \beta_{34}\, Year^3{}_{ij}\ , \ \ if\ Girls\ from\ Mixed\ Schools \end{cases} \tag{3.7}$$

Where,

β_{0k} = Intercepts for different school types and gender combinations.

β_{1k}= Linear slope parameters for each of the school type and gender combinations.

β_{2k} = Quadratic slope parameters for each of the school type and gender combinations.

β_{3k} = Cubic slope parameters for each of the school type and gender combinations.

3.4.2 The Marginal Mean Model

One of the key features in GEEs for analyzing longitudinal data is in ensuring we correctly specify how the mean of the outcome variable Y_{ij} is related to the covariates of our interest. We

define the mean structure of the outcome variable as,

$$\mu_{ij} = \pi_{ij}(x_{ij}) = P[Y_{ij} = 1 — x_{ij}] = E[Y_{ij} — x_{ij}] \tag{3.8}$$

μ_{ij} is the marginal expectation of the response which depends on the covariates x_{ij} through a known link function given by,

$$g(\mu_{ij}) = x'_{ij}\beta = \log it_{ij} \tag{3.9}$$

3.4.3 Specification of Working Covariance and Correlation Matrix.

Let the variance of the binary response variable Y_{ij} be denoted as $var(Y_{ij})$. The variance of each of Y_{ij} given the effects of the covariates, depends on the mean response.

$$Var(Y_{ij}) = \mu_{ij}(1 - \mu_{ij}) = f(\mu_{ij}) \tag{3.10}$$

We specify a correlation matrix $R_i(\alpha)$ such that it is close to the true correlation of the response. $R_i(\alpha)$ is a working correlation matrix and models the dependence between the within cluster observations.

The variance covariance matrix can be written as,

$$V_i = (A^{\frac{1}{2}}{}_i R_i(\alpha) A^{\frac{1}{2}}{}_i)\phi \tag{3.11}$$

where

ϕ is an overdispersion parameter

α is a vector of parameters describing the within-subject correlation

A_i is the diagonal matrix with marginal variances on the main diagonal.

$$A_i = \begin{pmatrix} \text{var}\left(y_{ij1|xij1}\right) & 0 & 0 & 0 & 0 \\ 0 & \text{var}\left(y_{ij2|xij2}\right) & 0 & 0 & 0 \\ 0 & 0 & \text{var}\left(y_{ij3|xij3}\right) & 0 & 0 \\ 0 & 0 & 0 & \text{var}\left(y_{ij4|xij4}\right) & 0 \\ 0 & 0 & 0 & 0 & \text{var}\left(y_{ij5|xij5}\right) \end{pmatrix} \tag{3.12}$$

V_i is known as the working covariance matrix of Y_i.

GEE models the correlation matrix by use of several correlation structures such as the independent correlation structure, exchangeable/compound symmetry correlation structure, AR (1) correlation structure and unstructured correlation structure.

This study utilizes exchangeable correlation structure which assumes constant correlations between any two measurements within a subject for all time periods.

$$Corr\left(Y_{ij}, Y_{ir}\right) = \alpha, 0 < \alpha < 1 \tag{3.13}$$

$$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i(n_i-1)}\sum_{j\neq r} e_{ij}e_{ir} \tag{3.14}$$

In this case, the exchangeable correlation between two observations is given by

$$Corr\left(Y_{ij}, Y_{ir}\right)\begin{bmatrix} 1 & \propto & \propto & \propto & \propto \\ \propto & 1 & \propto & \propto & \propto \\ \propto & \propto & 1 & \propto & \propto \\ \propto & \propto & \propto & 1 & \propto \\ \propto & \propto & \propto & \propto & 1 \end{bmatrix} \tag{3.15}$$

Important to note though is that GEE has the property of being a consistent estimator of the covariance matrix of β, even if the working correlation is mis-specified. (Lawal, 2003).

3.5 Parameter Estimation in GEE

The GEE estimator for for marginal models while accounting for correlation in longitudinal data arises from minimizing an objective function and solving a set of score equations iteratively until convergence is achieved.

3.5.1 Objective Function

Liu (2010) shows that the GEE estimator for β arises from minimizing the objective function shown below with respect to β to obtain a score equation.

$$\sum_{i=1}^{N} [y_i - \mu_i(\beta)]^T V_i^{-1} [y_i - \mu_i(\beta)] \tag{3.16}$$

where μ_i is a vector of mean responses with the elements $\mu_{ij} = \mu_{ij}(\beta) = g^{-1}(x_{ij}\prime\beta)$.

3.5.2 Score Equations/Estimating Function

The score function is obtained as a result of minimizing the above objective function , denoted as $S(\beta)$ and is of the form,

$$S(\beta) = \sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \beta_j} \left(A_i^{1/2} R_i(\alpha) A_i^{1/2} \right)^{-1} \phi (y_i - \mu_i) = 0 \tag{3.17}$$

Substituting $V_i = \left(A_i^{1/2} R_i(\alpha) A_i^{1/2} \right) \phi$ and $D^T = \frac{\partial \mu_i}{\partial \beta_j}$, we have the estimating equation whose solution is the GEE estimator of β;

$$S(\beta) = \sum_{i=1}^{N} D^T (V_i)^{-1} (y_i - \mu_i) = 0 \tag{3.18}$$

D^T is a Jacobian $n_i * p$ matrix given by;

$$D^T = \frac{\partial \mu_i}{\partial \beta_j} = \begin{pmatrix} \partial\mu_{i1}/\partial\beta_1 \; \partial\mu_{i1}/\partial\beta_2 \ldots\ldots\ldots \partial\mu_{i1}/\partial\beta_p \\ \partial\mu_{i2}/\partial\beta_1 \; \partial\mu_{i2}/\partial\beta_2 \ldots\ldots\ldots \partial\mu_{i2}/\partial\beta_p \\ . \\ . \\ \partial\mu_{in_i}/\partial\beta_1 \; \partial\mu_{in_i}/\partial\beta_2 \ldots\ldots\ldots \partial\mu_{in_i}/\partial\beta_p \end{pmatrix} \tag{3.19}$$

Where $(y_i - \mu_i)$ is a residual vector which measures deviations of observed responses of the i_{th} subject (school) from its mean.

This estimating equation is unbiased regardless of which covariance matrix V_i we use as long as we correctly defined the mean structure i.e $E[S(\beta)] = 0$.

3.6 Standard iterative procedure for GEE Parameter Estimation.

Parameter estimation in GEE is based on an algorithm for an iterative procedure in solving the score equation $S(\beta) = 0$, until the estimates obtained from the score equation converge.

3.6.1 Fisher Scoring.

The Fisher scoring method uses the expected derivative of the score, otherwise known as the Fishers information matrix. The procedure is as follows,

1. Compute initial estimates of for β; say $\hat{\beta}^{(0)}$, using univariate GLM i.e. assuming independence or rather using conventional logistics regression.

2. Given $\hat{\beta}^{(0)}$, compute method of moments estimates for α (if its unknown). With the obtained estimates for α, we compute $R_i(\alpha)$ and consequently the estimate of covariance

of $V_i = (A_i^{1/2} R_i(\propto) A_i^{1/2})\phi$

3. After t iterations we have say $\hat{\beta}^{(t)}$ and update the estimator for $\hat{\beta}$ by solving the estimating equation using the fishers scoring algorithm to obtain improved estimates:

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + \left(\sum_{i=1}^{N} D_i^T V_i^{-1} D_i \right)^{-1} \times \sum_{i=1}^{N} D^T V_i^{-1} (y_i - \mu_i) \tag{3.20}$$

4. Evaluate convergence using changes in $||\hat{\beta}^{(t+1)} - \hat{\beta}^{(t)}||$.

   We iterate the above procedure until convergence criterion is satisfied. Convergence occurs when there is no much improvement in the quasi likelihood value, or if the set threshold for the change in quasi likelihood is reached. Usually when the change is less than 0.0001 (SAS convergence tolerance).

3.6.2 Estimating α using method of moments

We recall that,

$$Var(Y_{ij}) = \mu_{ij}(1 - \mu_{ij}) = f(\mu_{ij}) \tag{3.21}$$

and obtain the Pearson residuals ϵ_{ij} using the moment based estimate as follows;

$$e_{ij}(\beta) = \frac{Y_{ij} - \hat{\mu}_{ij}(\hat{\beta})}{[\text{var}(Y_{ij})]^{1/2}} \tag{3.22}$$

The correlation parameter α is estimated as a simple function of ϵ_{ij} depending on the choice of correlation structure.

For exchangeable correlation structure, moment based estimator for α is given by,

$$\hat{\propto} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i(n_i - 1)} \sum_{j \neq r} e_{ij} e_{ir}; Corr(Y_{ij}, Y_{ir}) = \propto \tag{3.23}$$

For AR(1) correlation structure,

$$\hat{\propto} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i - 1}\sum_{j\leq n_i-1} e_{ij}e_{i,j+1}\ ;\ Corr(Y_{ij}, Y_{ir}) = \propto^{|j-r|} \tag{3.24}$$

For the unstructured correlation,

$$\hat{\propto}_{jr} = \frac{1}{N}\sum_{i=1}^{N} e_{ij}e_{ir}\ ;\ Corr\left(Y_{ij}, Y_{ir}\right) = \propto_{jr} \tag{3.25}$$

3.7 Goodness of Fit Statistics-QIC.

GEE method is based on the quasi likelihood theory and therefore the Akaikes Information Criterion (AIC), which is a widely used method for model selection in GLM, is not applicable to GEE directly. AIC computation requires a full conditional likelihood which is not obtained under GEE. However, a model-based selection method for GEE known as Quasi-likelihood under the Independence model Criterion, denoted (QIC) is largely used. QIC statistics allow for marginal model selection as well as selection of correlation structures through comparisons of fitted GEE models.

3.7.1 Quasi-likelihood under the Independence model Criterion-QIC.

QIC is basically an appropriate modification of the widely used Akaikes Information Criterion (AIC) to allow for model selection in GEE. The mathematical theory of QIC thus originates from the general formulation of AIC given by,

$$AIC = -2LL + 2p \tag{3.26}$$

where LL is the log likelihood and p is the number of parameters in the model.

QIC is derived by modifying the above formula and adjusting for the penalty term $2p$ as follows,

$$QIC = -2\sum_{i}\sum_{j} Q_{ij}\left(\hat{\mu}_{ij}; I\right) + 2trace(\hat{\Omega}_I^{-1}\hat{V}_R) \tag{3.27}$$

Where,

- I is the independent covariance structure used to calculate the quasi-likelihood.
- $\hat{\mu}_{ij} = g^{-1}(x'_{ij}\beta)$ and $g^{-1}(.)$ is the inverse link function.
- $\hat{V_R}$ is the robust variance estimator obtained from a general working covariance structure R.
- $\hat{\Omega}_I$ is another variance estimator obtained under the assumption of an independence correlation structure $\hat{\Omega}_I = \sum_{i=1}^{N} D_i^T V_i^{-1} D_i$.

Trace here refers to the sum of the diagonal elements of the matrix.

The quasi-likelihood in this model is of the general form;

$$Q_{ij} = \ y_{ij} ln\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) + \ln(1-\mu_{ij}) \tag{3.28}$$

A model with the smallest QIC value for a given correlation matrix is chosen as the preferred correlation structure.

A subset of covariates with the smallest QIC value is the preferred model.

Model selection and correlation structure will therefore be done in two stages.

1. First fix the mean structure and compare models with different covariance structures. The covariance structure with lowest QIC value is the best.
2. Subsequently, fix the covariance structure obtained in step 1 above and compare models with different mean structure. The model that yields the smallest QIC value is chosen as the best model.

3.8 Inference on β

3.8.1 Variance of β

GEE yields two versions of $Var\left(\hat{\beta}\right)$: the robust or empirical and model-based standard errors. The solution $\hat{\beta}$ is consistent and asymptotically normal.

Model based or Nave estimate.

The model based estimate for the variance of $\hat{\beta}$ assumes that the correlation model is correct and is obtained by,

$$\Sigma_M = M_0^{-1} = \sum_{i=1}^{N} D_i^T V_i^{-1} D_i \tag{3.29}$$

This is usually a GEE equivalent of the inverse of the Fisher information matrix which is often used in GLMs as an estimator of covariance estimate of the MLE of β.

Robust /Sandwich estimator.

The sandwich estimator, also known as robust or empirical accounts for a correlation model that is not correct and is obtained by $\Sigma_R = M_0^{-1} C M_0^{-1}$ where ,

$$\begin{aligned} C &= \sum_{i=1}^{N} D_i^T V_i^{-1} (y - \hat{\mu})(y - \hat{\mu})^T V_i^{-1} D_i \\ C &= \sum_{i=1}^{N} D_i^T V_i^{-1} \widehat{Cov}(Y_i) V_i^{-1} D_i \end{aligned} \tag{3.30}$$

Σ_R is called the empirical, robust or sandwich variance estimate.

If $C = (y - \hat{\mu})(y - \hat{\mu})^T$, then the model based estimate is equal to the sandwich estimator since $C \approx M_0$. This will occur only if the true correlation structure is correctly modelled. Otherwise $\Sigma_R \neq \Sigma_M$.

One of the properties of this estimator is that it provides a consistent estimator of $V\left(\hat{\beta}\right)$ even if the working correlation structure is not the true correlation of Y.

Generalized Score Statistics

In GEE, score tests are used in testing the hypothesis Ł$\beta = 0$, where L is usually a user-specified $c * d$ matrix or a contrast for Type 3 test of hypothesis.

Given $\tilde{\beta}$ is a regression parameter obtained from solving GEE under the restricted model Ł$\beta = 0$, and $S(\beta)$ as the generalized estimating equation values at $\tilde{\beta}$, the generalized score statistic is given by;

$$T = S\Big(\tilde{\beta}\Big)' \Sigma_M L' \big(L \Sigma_R L'\big)^{-1} L \Sigma_M S(\tilde{\beta}) \tag{3.31}$$

Where,

Σ_M is the Model-Based covariance estimate,
Σ_R , is the Robust/Empirical covariance estimate.

The p-values for the generalized score statistic are computed based on the chi-square distribution with c degrees of freedom.

Chapter 4

DATA ANALYSIS AND FINDINGS

4.1 Introduction

In our analysis, we utilize two statistical softwares namely SAS and R 3.1.2 in combination with R-Studio. We use R statistical software in performing necessary data manipulations as well as performing exploratory data analysis. We further use SAS software in performing GEE analysis as well as performing contrasts of interest. In this study, we analyze the data at a significance level $\alpha = 0.05$.

Some of the key packages we use in R-Gui include ggplot2, gridextra, plyr, reshape, and xtable. We further use the SAS procedure PROC GENMOD in fitting the model and present the GEE results.

4.2 Exploratory data analysis (EDA).

Exploratory data analysis, denoted EDA, usually focuses on exploring the data so that one understands the variables and data structure, and thus develops an intuition about the data set. It provides a summary of the data under study. In this section we perform EDA in R statistical software and present necessary summaries that provide basic information about the data.

4.2.1 Data summary

The study covered 237 unique schools within Nakuru district for the 5 year period. It was noted that the number of schools have been declining over the years. Caution should be taken however in making such a conclusion as it is possible that the actual number of schools did not reduce, but the reporting of results on the Kenya open data website was the one that was not efficiently done. Below is a summary of the number of schools under study within the 5 year period, for different school types.

Table 4.1: Data summary

	Year				
School type	2006	2007	2008	2009	2010
Girls Only	26	28	7	5	5
Boys Only	14	14	6	7	4
Mixed	157	176	55	60	34
Total	197	218	68	72	43

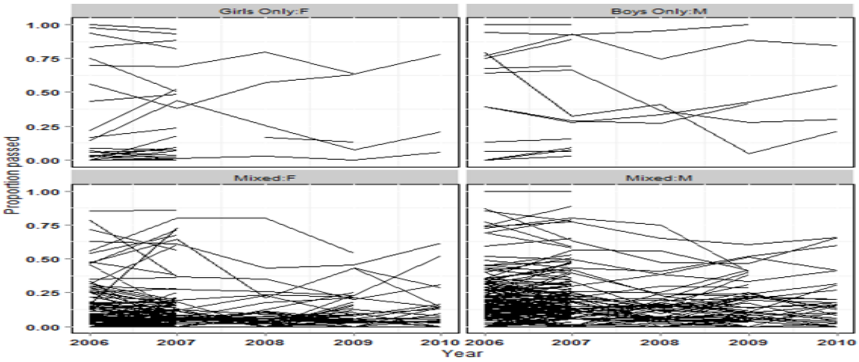
Subject specific profiles

Below is a graphical representation for the subject specific profiles of the proportion of students that passed over the five year period. The left column panels are for girls while the right column panels represent the boys. From the plots, it is clear that there are fewer schools whose results are available in later years. This missingness pattern is however not analyzed in this thesis and data is assumed to be missing at random.

```
>nakuru.wide$School.type.Gender <- nakuru.wide$School.type:nakuru.wide$Gender
 #Interaction between school type and gender
>ggplot(data=nakuru.wide, aes(x=Year,y=prop.pass ,group=KNEC.Code:Gender))
+ylab('Proportion passed')+geom_line()+facet_wrap(~School.type.Gender)+
theme(legend.position=c(0.75,0.75),panel.background=element_rect(fill='white',
colour='black'),axis.text.x=element_text(face='bold',colour='black')
```

```
,axis.text.y=element_text(face='bold',colour='black'))
```

Figure 4.1: Subject specific profiles for proportion of students that passed over 5 year period.

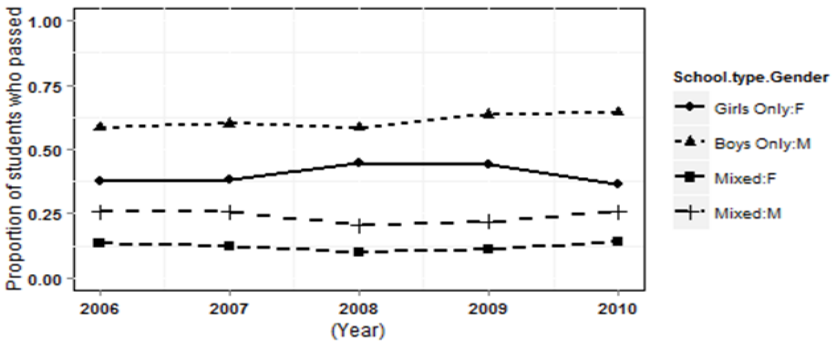


Average profiles

Figure 2 below is a visual of the average profiles for the proportion of students that passed over the five year period. Its observed that on average, boys from boys schools have a higher pass rate over the years. Girls from mixed schools have a lower pass rate compared to boys from mixed schools. Moreover, for a given school type and gender, the pass-rate seems relatively constant over the years.

```
>ggplot(data=avg.nakuru.wide,aes(x=  (Year),y=prop.pass ,group=School.type.Gender,
 shape=School.type.Gender,linetype=School.type.Gender))+geom_point(size=3)
 +ylim(0,1) +ylab('Proportion of students who passed')+geom_line(size=1)
 +theme(panel.background=element_rect(fill='white',colour='black'),
 axis.text.x=element_text(face='bold',colour='black'),
 axis.text.y=element_text(face='bold',colour='black'))
```

Figure 4.2: Average profiles for proportion of students that passed over 5 year period.



4.3 GEE analysis

GEE analysis was performed using the SAS procedure PROC GENMOD. We fitted various mean models and incorporated different working correlation matrices for the covariance structure until we identified the best fit based on QIC values. For ease of model convergence, we centered the years with 2006 as the base year.

4.3.1 Model Fitting

In fitting the models, we did not include the usual common intercept as we were not interested in its interpretation. We fitted two different mean models as below and the best fitting model was the one corresponding to smallest QIC value.

Model with school, gender specific intercepts but shared slopes

$$Model1: \ Y_i = \beta_{0k} schoolsex + \beta_{1k} yearr_i + \beta_{2k} year_i^2 + \beta_{3k} year_i^3 \tag{4.1}$$

Model with school, gender specific intercepts and school/gender specific slopes.

$$Model2: \ Y_i = \beta_{0k} schoolsex + \beta_{1k} schoolsex * year_i + \beta_{2k} schoolsex * year_i^2 + \beta_{3k} schoolsex * year_i^3 \tag{4.2}$$

4.3.2 Selection of working covariance structure

The selection of a covariance structure for this study was based on the four working correlation matrices, i.e. Independent, Exchangeable/Compound symmetry, Unstructured and Auto-Regressive (AR1). We first fixed *Model 1* as the mean structure and adjusted for the working correlation matrix while comparing the QIC values from model output.

4.3.3 Goodness of fit statistics.

We used the quasi-likelihood under independence model criterion statistics (QIC) to select the best fitting model as well as the working correlation for the covariance structure. Table 2 is a summary for the QIC values obtained from fitting the 2 models above and different working correlation matrices as earlier defined. *Model 1* with the school specific intercepts and shared

Table 4.2: Goodness of fit statistics

SAS GEE Fit Criteria-QIC Values				
Label	Independence	Exchangeable	Auto-regressive (AR1)	Unstructured
Model 1	172.1338	**142.427**	142.8863	0.0000a
Model 2	192.343	146.8489	149.6195	0.0000a

a: Model did not converge hence no reported value

slopes was selected for this study as it had the smallest QIC values. None of the models converged under the unstructured working correlation.

Based on the QIC statistics above, we further selected a covariance structure per school with an exchangeable/compound symmetry working correlation matrix. This implies that the correlation is shared between boys and girls over the 5 years regardless of the school type. We fitted this model and performed further inference from the results of this model.

4.3.4 GEE Results

In this section we present GEE results based on *model1* and a covariance structure with an exchangeable correlation matrix. The correlation between measurements of the same school was obtained as 0.837 which is very high an indication that the measurements were highly correlated hence the need to account for clustering.

Table 4.3: Working correlation for fitted model

Exchangeable Working Correlation	
Correlation	0.837

Score Statistics

The overall significance test based on a score test is presented in Table 4. The score chi-square statistic is computed based on the generalized score function. In GEE, type 3 analysis uses the

Table 4.4: Score Statistics

Score Statistics For Type 3 GEE Analysis			
Source	DF	Chi-Square	Pr $\geq$ ChiSq
School sex	4	87.82	$\leq .0001$
$Year$	1	1.18	0.2783
$Year^2$	1	3.04	0.0811
$Year^3$	1	4.03	0.0447

likelihood ratios instead of the usual sum of squares by defining an estimable function for an effect of interest. The score statistics indicate that there is a significant difference between the intercepts. Thus, the hypothesis for equal pass rates for different school types and gender in 2006 is rejected, implying that the performance of the schools differed with 2006 as the base year.

$$\begin{aligned} H_0 &: \ \beta_{01} = \beta_{02} = \beta_{03} = \beta_{04} \\ H_1 &: \ \beta_{0i} \neq \beta_{0j} \end{aligned} \tag{4.3}$$

On the other hand, there is no significant effect of the time. This means that the pass-rate in KCSE does not depend on the number of years elapsed since 2006. This is consistent with our intuition in the exploratory average plot where we concluded that there were no significant changes in the slopes over time.

Parameter Estimates

Parameter estimates for the model coefficients are as presented in Table 5. Both model-based and empirical standard errors are shown. Empirical standard errors are observed to be generally larger than model based standard errors. This can generally be attributed to the fact that with highly correlated data, there are fewer observations contributing to independent information as compared to the case of model-based estimation which assumes the dataset is truly independent. Thus the effective sample size resulting from GEE with exchangeable correlation is given by $N_{eff} = \sum_{i=1}^{N} \tilde{n}_i = \sum_{i=1}^{N} \frac{n_i}{1+\rho(n_i-1)}$ where N_{eff} is the effective sample size corresponding to the truly independent samples in this correlated data, n_i is the number of repeated measurements per subject and ρ is the correlation coefficient as estimated from the GEE estimation. It is easy to see that if the correlation is zero, then GEE analysis provides information using a sample size similar to that from independent data.

Table 4.5: Parameter estimates

SAS Analysis Of GEE Parameter Estimates					
	Estimate	Model based SE	95% CI	Empirical based SE	95% CI
Boys in Boys only school	0.146	0.1758	(-0.199,0.491)	0.3951	(-0.628,0.92)
Girls in Girls only school	-0.7746	0.1555	(-1.079,-0.47)	0.3421	(-1.445,-0.104)
Boys in Mixed schools	-1.2728	0.0872	(-1.444,-1.102)	0.1217	(-1.511,-1.034)
Girls in Mixed schools	-2.1455	0.1347	(-2.41,-1.882)	0.172	(-2.483,-1.808)
$Year$	0.102	0.0874	(-0.069,0.273)	0.0967	(-0.088,0.292)
$Year^2$	-0.1284	0.0628	(-0.252,-0.005)	0.0723	(-0.27,0.013)
$Year^3$	0.0265	0.0112	(0.005,0.049)	0.0124	(0.002,0.051)

Contrast Estimate Results.

The test of hypothesis of interest now reduces to the test of whether there were differences in performance across different gender between mixed schools and single sex schools. Thus we perform contrasts tests for the intercepts only. To achieve this, the 'ESTIMATE' statement was used in SAS. Results are presented in the Table 6.

Table 4.6: Contrast estimates

Contrast Estimate Results						
Contrast Label	Mean Estimate	95% CI	L'Beta Estimate	95% CI	Chi-Square	Pr $\geq$ ChiSq
Boys only vs Boys mixed	0.8052	(0.6444,0.904)	1.4188	(0.5947, 2.243)	11.38	0.0007
Girls only vs Girls mixed	0.7975	(0.654,0.8914)	1.3709	(0.6365,2.1053)	13.38	0.0003
Boys mixed vs Girls mixed	0.7053	(0.6773,0.7318)	0.8726	(0.7413,1.004)	169.6	¡.0001

The LBeta column represents the difference in parameter estimates (log (OR)) that were shown in (Table 5: Parameter estimates). For instance, for the hypothesis on the difference between boys in boys only school versus boys in mixed schools, the LBeta estimate is given by;

$$L'Beta = (\beta_{i1} - \beta_{i3}) = \{0.146 - (-1.2728)\} = 1.4188 \tag{4.4}$$

The mean estimate column denotes the probability of success for the contrast under review. Thus for the above case on boys in boys school only versus boys in mixed schools, the mean estimate is given by;

$$\begin{aligned} MeanEstimate &= \frac{\exp\{\log(OR)\}}{1+\exp\{\log(OR)\}} = \frac{exp(\beta_{i1}-\beta_{i3})}{1+exp(\beta_{i1}-\beta_{i3})} = \frac{\exp(1.4188)}{1+\exp(1.4188)} \\ &= \frac{4.1322}{1+4.1322} = 0.8052 \end{aligned} \tag{4.5}$$

The results indicate that there was a significant difference between performances of boys in boys only schools versus boys in mixed schools. Boys in boys schools only had an 80.52% probability of passing compared to boys in mixed schools.

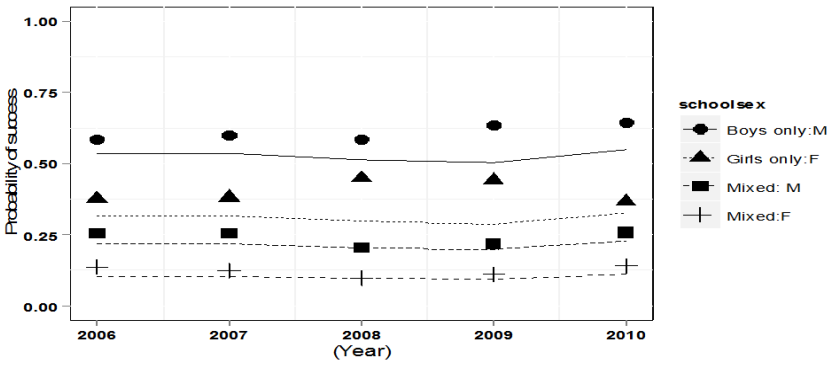
Similarly, there was a significant difference between performances of girls in girls school only versus girls in mixed schools. Girls in a girls only school had a 79.75% probability of passing

KCSE compared to girls in a mixed school.

Finally, there was a significant difference in KCSE performance for boys in mixed schools compared to girls in mixed schools. Boys in mixed schools had a 70.53% probability of passing compared to girls in mixed schools.

To conclude, a graphical presentation of the observed and predicted values of the average pass rate is presented in Figure 3. The model fits nicely data for mixed schools although for boys only and girls only schools, there is some variability in the slope components.

Figure 4.3: Average pass rate for observed vs predicted.



Chapter 5

DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS

5.1 Discussion.

The study aim was to gain insights on KCSE performance in Nakuru County while focusing on the relationship between students gender as well as school type relative to their performance over time. A generalized estimating equations analysis was performed on longitudinal data for KCSE performance for the period 2006-2010 to account for possible correlations in performance of a school over time. Results from the analysis exhibited constant correlations (Exchangeable) in performance of schools over time.

The analysis further revealed significant differences in KCSE performance for single sex schools and mixed schools. Contrasts were performed to access one gender student performance in single sex schools against same gender in mixed schools. Results showed significant differences in performance with student from single sex schools having a higher pass rate than those in mixed schools. This is consistent with previous studies conducted by Mburu (2013) in Kericho and Kipkelion districts where he tried to establish if social classroom interaction had an effect on male and female student academic performance. The results are also consistent with those from

a report by Charnley (2008) in accessing GCSE performance of independent pupils based on gender and school type differences, where he showed that pupils from single sex schools performed significantly better in most subjects compared to their counterparts in mixed schools.

5.2 Conclusion.

In conclusion, there is evidence that students of a particular gender in one gender school perform better than they would in mixed schools. Moreover, girls in mixed schools are more disadvantaged as is evident from the low pass rate compared to boys in mixed schools. These conclusions are independent of the year under review since the slope components were not significant. Thus regardless of the year under review, male/female students in one-gender school perform better than males/females respectively in mixed gender schools.

5.3 Recommendations.

Having established that significant differences exist between student performance in KCSE amongst the single sex schools and mixed schools, its imperative that the ministry of education as well as other relevant education stakeholders formulate education policies geared towards an improved performance especially in mixed schools. The study especially strongly recommends keeping a closer look at the girl child in mixed schools by addressing arising distractions that are a hindrance to better performance.

5.4 Suggestion for further studies.

Further studies should focus on establishing factors associated with differences in KCSE performance in different school types as well as students gender.

Chapter 6

REFERENCES

Bagaka's, J. G. (2011). The role of teacher characteristics and practices on upper secondary school students' mathematics self-efficacy in nyanza province of kenya: A multilevel analysis. *International Journal of Science and Mathematics Education*, 9(4):817–842.

Charnley, J. (2008). The gcse performance of independent schoolpupils: gender and school type differences. Technical report, Center for Evaluation and Monitoring.

Eisenkopf, G., Hessami, Z., Fischbacher, U., and Ursprung, H. (2011). Academic performance and single-sex schooling: evidence from a natural experiment in switzerland. Technical report, University of Konstanz.

Ghisletta, P. and Spini, D. (2004). An introduction to generalized estimating equations and an application to assess selectivity effects in a longitudinal study on very old individuals. *Journal of Educational and Behavioral Statistics*, 29(4):421–437.

Lawal, B. (2003). *Categorical data analysis with SAS and SPSS applications*, chapter Analysis of repeated measures data, pages 506–534. Lawrence Erlbaum associates.

Liu, Z. Z. (2010). Marginal models: Generalized estimating equations.

Makewa, L. N., Role, E., and Ngila, W. M. (2014). Girl child challenges and academic achieve-

ment in mixed secondary schools. *Journal of Education and Human Development*, 3(2):471–491.

Mburu, D. D. N. P. (2013). Effects of the type of school attended on students academic performance in kericho and kipkelion districts, kenya. *International Journal of Humanities and Social Science*, 3(4):79–89.

Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data.* Springer series in statistics. Springer.

Njoroge, J. K. and Kerei, K. O. (2012). Free day secondary schooling in kenya: An audit from cost perspective. *International Journal of Current Research*, 4(3):160–163.

Weaver, M. A. (2009). Introduction to analysis methods for longitudinal/clustered data, part 3: Generalized estimating equations. *The International Clinical Studies Support Center (ICSSC).*

Yara, P. O. and Catherine, W. W. (2011). Performance determinants of kenya certificate of secondary education (kcse) in mathematics of secondary schools in nyamaiya division, kenya. *Asian Social Science*, 7(2):107–112.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44(4):1049–1060.

Zorn, C. J. W. (2001). Generalized estimating equations model for correlated data:a review with applications. *American Journal of Political Science*, 45(2):477–490.

Chapter 7

APPENDICES.

7.1 R Codes for data manipulations

```
>Nakuru <- kcse [which(kcse$County==’Nakuru’),]
>schoolmix <- ddply(Nakuru,.(KNEC.Code,Year),transform,
School.type=sum(as.numeric(unique(Gender))))

>schoolmix <- schoolmix [order(schoolmix$KNEC.Code),] #To sort by KNEC.Code

>schoolmix$School.type <- factor (schoolmix$School.type, levels=c (1:3),
labels= c (’Girls Only, Boys Only’,’Mixed’)) #Label the different school types

>save (schoolmix, file=’kcse.Rda’) #Save data in R data file for further analysis
> kcse <-   get(load(’kcse.Rda’))  #Load previously saved data
>kcse$school.gender <- as.factor(kcse$School.type:kcse$Gender)
 #Interaction between school type and gender
>unique(kcse$school.gender) #Check if results are ok

> as.character(unique(kcse$District.Name))
```

```
>nakuru <- kcse[which(kcse$District.Name==’NAKURU’),] #Confirm only Nakuru data in use

>unique.schooltype <- nakuru[!duplicated(nakuru[,c(’KNEC.Code’, ’School.type’)]),]
 # Subset unique records per school
>(all.duplicates <- unique.schooltype[duplicated(unique.schooltype[,’KNEC.Code’]),])
#Identify schools which have multiple school types

>(with.duplicates <-  (as.character(unique.schooltype[duplicated(unique.schooltype
[,’KNEC.Code’]),’KNEC.Code’])))
##Get unique KNEC>Code identifiers for schools with duplicate school types

>nakuru <- nakuru[which(!as.character(nakuru$KNEC.Code)%in%with.duplicates),]
#Remove duplicate school type records from the Nakuru data

>nakuru$pass <-   ifelse((as.character(nakuru$Grade.attained)==’A’
|as.character(nakuru$Grade.attained)==’A-’|as.character(nakuru$Grade.attained)==’B+’
|as.character(nakuru$Grade.attained)==’B’|as.character(nakuru$Grade.attained)==’B-’
|as.character(nakuru$Grade.attained)==’C+’),1,0)

>nakuru <- ddply(nakuru,.(Year,KNEC.Code, District.Name,
 School.Code,School.Name,pass,School.type,Gender)
,summarize,no.pass=sum(Frequency))
 #Count the number of students who passed/failed per school, gender and year

>nakuru.wide <- data.frame(reshape(nakuru,timevar = "pass",idvar = c("Year", "KNEC.Code",
"Gender", "School.type","District.Name", "School.Code", "School.Name"),
direction = "wide"))
```

```
>names(nakuru.wide)[names(nakuru.wide)%in%c('no.pass.0','no.pass.1')]
<- c('failed','passed')
```

7.2 SAS codes for GEE analysis

```
PROC IMPORT OUT= WORK.nakuru
            DATAFILE= "E:\Documents\Project Data\Analysis\nakuru_knec_wide.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;
proc format;
value schoolsex 1='Boys only:M' 2='Girls only:F' 3='Mixed:F' 4='Mixed: M';
run;

data nakuru;set nakuru;
yearcen=year-2006;
yearcensq=yearcen*yearcen;
yearcencube=yearcensq*yearcen;
total=passed+failed;

schoolsex=5;
if school_type='Boys Only' and gender='M' then schoolsex=1;
if school_type='Girls Only' and gender='F' then schoolsex=2;
if school_type='Mixed' and gender='F' then schoolsex=3;
if school_type='Mixed' and gender='M' then schoolsex=4;
format schoolsex schoolsex.;
run;
```

```
*Final GEE model: compound symmetry/exchangeable;
proc genmod data=nakuru ;
class school_type gender knec_code schoolsex year subjectid;
model passed/total=schoolsex yearcen yearcensq yearcencube
/ dist=bin link=logit noint p type3 ;
repeated    subject=knec_code/ type=cs   corrw modelse  ;
 ods output obstats=predmodel;
 *Test the hypotheses:;
*performance was the same in the beginning (2006)-
 test if intercepts are equal between schooltype/gender;
estimate 'Int: Boys only vs Boys mixed' schoolsex 1 0 -1  0/e;
estimate 'Int: Girls only vs Girls mixed' schoolsex 0 1 0 -1/e;
estimate 'Int: Boys mixed  vs Girls mixed' schoolsex 0 0 1 -1/e;
run;quit;
goptions reset=all ;
 proc gplot data=predmean;
 plot pred*year=schoolsex;
 plot2 obsprop*year=schoolsex;
 symbol i=join;
 symbol2 i=dot;
run;quit;
```