Analysis of KCSE performance in Nakuru county: A generalized estimating equations approach.

Elvis Muchene
University of Nairobi, Kenya
Nelson Owuor
University of Nairobi, Kenya

In the Kenyan education system, progression to 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 quasilikelihood under independence model criterion (QIC) was used to select best model as well as best 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.

Keywords : *Generalized estimating equations, KCSE, QIC.*

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.

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.

Every year around February-March, a ceremony headed by the minister of education is held where results of the KCSE examination of the preceding year are released. 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.

In this article,we assess the temporal effect of the reported performance of schools between the year 2006 and 2010 in Nakuru county. One of the challenges with this is that the scores from a school for the five years under review are correlated. Considering the outcome to be binary in nature, (C+ or higher=Pass, Less than C+ = Fail), marginal models such as the so called Generalized Estimating Equations (GEE) of Zeger et al. (1988) as well as the genelarized Linear Mixed Models (GLMM) of Molenberghs and Verbeke (2005) provides a mechanism to account for this correlation in the binary outcome.

The main aim of this study is to gain insights on KCSE performances in Nakuru County while accounting for 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 education stakeholders.

Specifically, the study aims to;

- Establish if indeed there exists a significant difference in overall KCSE performance between mixed schools, Boys Schools & Girls schools in Nakuru County
- Establish if boys performance differs significantly between mixed schools and boys schools
- Establish if girls performance differs significantly between mixed schools and girls schools
- Establish if there exists a significant difference in overall KCSE performance between boys & girls in mixed school.

This paper is structured as follows; In section 2, an overview of the dataset used in this analysis is presented as well as a detailed discussion of the statistical methodology, i.e GEE concept We present the results of the GEE analysis in section 3, and finally in section 4 we give a brief discussion as well as conclusions in section 5.

2. Case study data

Longitudinal data on Kenya Certificate of Secondary Education (KCSE) performance was obtained from the Kenyan government open data website for the period 2006-2010 https://www.opendata.go.ke/Education/KCSE-Exam-Results-2006-to-2010/ycfy-7tnf.

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.

Although the well established schools had registered candidates for KCSE every year, the database provided by the Kenyan government has a lot of missing data. This in fact is the reason that the temporal relationship is only investigated for 2006-2010 since its the only publicly available dataset. There was a high rate of incompelete datasets that spanned across various school types. While there are methodologies to account for different mechanisms of missng data, in this article, we do not investigate the missingness with a formal statistical approach. Instead, we assume data is missing at random (MAR).

Data is considered missing at random (MAR) when the probability of missingness is independent of the vectore of missing outcomes, but may be dependent on observed outcomes. Valid inferences for GEE estimators can be obtained from data which is MAR or (MCAR) i.e. missing completely at random Zorn (2001).

Table 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

3. Methodology

3.1. *Marginal model for correlated data*

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.

3.1.1 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 (Lawal, 2003). 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. Model intercept corre-

sponds to probability of success in the year 2006.
Moreover, transformation of the centered 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.

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

This equation can further be decomposed into a $p * 1$ vector of covariates (X_{ij}) as below;

$$Y_i = \begin{cases} \text{Boys only school:} \\ \beta_{01} + \beta_{11} year + \beta_{21} Year^2 + \beta_{31} Year^3 \\ \text{Girls only school:} \\ \beta_{02} + \beta_{12} year + \beta_{22} Year^2 + \beta_{32} Year^3 \\ \text{Boys from mixed school:} \\ \beta_{03} + \beta_{13} year + \beta_{23} Year^2 + \beta_{33} Year^3 \\ \text{Girls from mixed school:} \\ \beta_{04} + \beta_{14} year + \beta_{24} Year^2 + \beta_{34} Year^3 \end{cases} \tag{2}$$

Where,

β_{ok}= 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.

3.1.2 Specification of working covariance and correlation matrix

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 (Molenberghs and Verbeke, 2005).

We denote the working covariance matrix of Y_i as V_i .This variance function of the observations within clusters is modeled by

$$V_i = (A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}})\phi \tag{3}$$

Where,
α Is a vector of parameters describing the within-subject correlation
ϕ is an over-dispersion parameter and,
$R_i(\alpha)$ Is the working correlation matrix to model the dependence between within cluster observations.

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.

3.2. *Parameter estimation in GEE*

The GEE estimator for β while accounting for correlation in longitudinal data arises from minimizing an objective function;

$$\sum_{i=1}^{N} [y_i - \mu_i(\beta)]^T V_i^{-1} [y_i - \mu_i(\beta)] \tag{4}$$

to obtain a set of score equations given by.

$$S(\beta) = \sum_{i=1}^{N} D^T (V_i)^{-1} (y_i - \mu_i) = 0 \qquad (5)$$

and the solution is the GEE estimator of β. 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 (β)] = 0

3.3. *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(β)=0, until the estimates obtained from the score equation converge.

3.3.1 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 it is unknown), compute $R_i(\alpha)$ and consequently the estimate of covariance of V_i= $(A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}})\phi$
3. At t iterations, update the estimator for β by solving the estimating equation using the fishers scoring algorithm to obtain improved estimates:

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + \left(\sum_{i=1}^{N} D_i^T V_i^{-1} D_i \right)^{-1} \times \sum_{i=1}^{N} D^T V_i^{-1} (y_i - \mu_i) \qquad (6)$$

We iterate the above procedure until convergence criterion is satisfied (Weaver, 2009). 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).

3.4. *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.

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.

QIC is derived by modifying the AIC formula and is given by

$$QIC = -2 \sum_i \sum_j Q_{ij} \left(\hat{\mu}_{ij}; I \right) + 2trace(\hat{\Omega}_I^{-1} \hat{V}_R) \qquad (7)$$

Where;

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

Model selection & correlation structure is 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

3.5. Parameter Inference.

3.5.1 Model based or Naive estimate.

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

$$\Sigma_M = M_0^{-1} = \sum_{i=1}^{N} D_i^T V_i^{-1} D_i \qquad (8)$$

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 $\hat{\beta}$

3.5.2 Robust /Sandwich estimator.

The sandwich estimator, also known as robust or empirical accounts for a correlation model that is not correct and is given by

$$\Sigma_R = M_0^{-1} C M_0^{-1} \qquad (9)$$

where,
$C = \sum_{i=1}^{N} D_i^T V_i^{-1} (y - \hat{\mu})(y - \hat{\mu})^T V_i^{-1} D_i$
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.

3.6. Generalized Score Statistics

In GEE, score tests are used in testing the hypothesis *Lβ=0*, where L is usually a user-specified $c * d$ matrix or a contrast for Type 3 test of hypothesis. The generalized score statistic is given by,

$$T = S(\tilde{\beta})' \Sigma_M L' \left(L \Sigma_R L'\right)^{-1} L \Sigma_M S(\tilde{\beta}) \qquad (10)$$

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.

4. Results

GEE analysis was performed using the SAS procedure PROC GENMOD. Various mean models were fitted and incorporated different working correlation matrices for the covariance structure until the best fit based on QIC values was identified.

4.1. 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.

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 as per the data summary table 1.

Caution should be taken however in making such a conclusion since it is possible that the actual number of schools did not reduce, but the reporting of results on the Kenya open data website was not efficiently done.

4.2. Goodness of fit statistics

Table 2 shows results for the goodness of fit statistics for two fitted models and various correlation structures. A model with the school specific intercepts and shared slopes was selected for this study since it had the smallest QIC values.

Table 2: Goodness of fit statistics. Ind: Independence, EXCH: exchangeable, AR1: Autoregressive

Label	Ind	CS	AR1
Model 1	172.1338	**142.427**	142.8863
Model 2	192.343	146.8489	149.6195

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. None of the models converged under the unstructured working correlation.

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.

4.3. *Score Statistics*

The overall significance tests based on a score test are presented in Table 3. The score chi-square statistic is computed based on the generalized score function.

Table 3: Score Statistics

Source	DF	Chi-Square	P-value
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

The score statistics show 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. On the other hand, there is no significant effect of the time. This means that the pass-rate in KCSE did not depend on the number of years elapsed since 2006.

4.4. *Parameter estimates*

Results for model coefficients parameter estimates are presented in Table 4. 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.

Table 4: Parameter estimates

	Estimate	Model based SE	95% CI
Boys in Boys only school	0.146	0.1758	(-0.199,0.491)
Girls in Girls only school	-0.7746	0.1555	(-1.079,-0.47)
Boys in Mixed schools	-1.2728	0.0872	(-1.444,-1.102)
Girls in Mixed schools	-2.1455	0.1347	(-2.41,-1.882)
Year	0.102	0.0874	(-0.069,0.273)
Year2	-0.1284	0.0628	(-0.252,-0.005)
Year3	0.0265	0.0112	(0.005,0.049)

4.5. *Contrast estimates*

The test of hypothesis of interest now reduced to the test of whether there were differences in performance across different gender between mixed schools and single sex schools. Thus we performed contrasts tests for the intercepts only. To achieve this, the *Estimate* statement was used in SAS. Results are presented in the Table 5.

Table 5: Contrast estimates

		Contrast Estimate Results		
Contrast Label	Mean Estimate	95% CI	L'Beta Estimate	95% C
Boys only vs Boys mixed	0.8052	(0.6444,0.904)	1.4188	(0.5947
Girls only vs Girls mixed	0.7975	(0.654,0.8914)	1.3709	(0.6365
Boys mixed vs Girls mixed	0.7053	(0.6773,0.7318)	0.8726	(0.7413

The $L - Beta$ column represents the difference in parameter estimates (log (OR)) that were shown in (Table 4: Parameter estimates). For instance, for the hypothesis on the difference between boys in boys only school versus boys in mixed schools, the $L - Beta$ estimate is given by;

$$L'Beta = (\beta_{i1} - \beta_{i3}) = \{0.146 - (-1.2728)\} \\ = 1.4188 \tag{11}$$

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;

$$MeanEstimate = \frac{\exp\{\log(OR)\}}{1+\exp\{\log(OR)\}} = \\ \frac{exp(\beta_{i1}-\beta_{i3})}{1+exp(\beta_{i1}-\beta_{i3})} = \frac{\exp(1.4188)}{1+\exp(1.4188)} \\ = \frac{4.1322}{1+4.1322} = 0.8052 \tag{12}$$

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.

5. Conclusion

This paper 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.

5.1. *Conclusion*

In conclusion, there is evidence that students of a particular gender in one gender school perform better than they would in mixed schools.

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.

5.2. *Recommendations*

Having established that significant differences exist between student performance in KCSE amongst the single sex schools and mixed schools, it is 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.

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

References

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

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

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.

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

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.

Correspondence: elviskaranja@gmail.com.