Applied Mathematical Sciences, Vol. 8, 2014, no. 68, 3359 - 3374
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ams.2014.44277

Application of Generalized Estimating Equation (GEE) Model on Students' Academic Performance

Isaac Owusu-Darko

Mathematics Department
Valley View University, Techiman,
P. O. Box 183, Techiman Ghana, W/A

Isaac Kwasi Adu

Mathematics Department
Valley View University, Techiman
P. O. Box 183, Techiman, Ghana, W/A

Nana Kena Frempong

Mathematics Department
Kwame Nkrumah University of Science and Technology
Ghana, West Africa

Abstract

This paper illustrates analysis of longitudinal data on students' academic performance using GEE (Generalize Estimation Equations) under various working correlation assumptions. Many factors account for students' academic performance in the fulcrum of all levels of education. Hence, any variable that triggers the academic performance of students evoke the awareness of all. The aim of this thesis is to analyze academic performance using application of Generalized Estimating Equation (GEE) Models under various working correlation assumptions. There are various statistical and mathematical models employed in the analyses of students' academic performance in different level of schools. In this paper, we formulate the Generalized Estimating Equation (GEE) model approach under various correlation assumptions to analyze the probable

performance in relation to variables such as gender, entry age into the school, the geographical location of students, as well as Graded level of former School attended. We used real data set of students' Semester Weighted Average (SWA), and back these with validate and reliable questionnaire about students personal information (on their Biodata response) for a complete data set.

From our analyses, the test of model-based and empirical-based standard error estimates on Coefficient Estimation of the Study Parameters based on our GEE assumptions reveals that, only the geographical location of students is significant and hence affects their academic performance. Our contrast effect of linear time interactions with respect to the locations made us recommend an enhancement of mathematics teaching in locations that are lagging, especially in the Northern Belt of Ghana.

Keywords: Generalized Estimating Equation model (GEE), Semester Weighted Average (SWA), Academic Performance (AP), Working correlation Assumptions, Geographical Location

1 Introduction

The academic performance of students in every institution is the concern of all and sundry; especially parents, stakeholders, teachers/lecturers, the government and among others. Due to this conceptual viewpoint, any constraining variable that might affect the performance of these students in tertiary institutions throws a concern of which people are interested to find the cause and effect of the issue.
Many reasons have been attributed for the rate of academic performance (whether negative or positive trend) in our tertiary institutions. Some people trace the cause rate to student inability to comprehend the principles of Mathematics. Others are of the view that the abysmal performance is due to loaded curriculum (there is too much to be taught within a short time) and among others. There is the need to explore our investigation to find out the extent at which socio-demographic factors (such as age, gender), geographical location of students, and former school attended also have effect on students academic performance. The peculiar nature of mathematics and the rate at which these factors could affect students' SWA scores have led to the research on the analysis of the performance of University students in KNUST Mathematics Department in relation to students' gender, age, geographical background location and graded level of former school attended. The study however sought to find out whether these socio-demographic factors also affect students Academic performance in their SWA scores.
The paper first presents the purpose of the paper and discusses the method used. It then gives a brief background of GEE models and the three measures used for model comparison. The paper concludes with a discussion of the estimation results and its recommendation for future research.
Generalized estimating equations (GEE) were introduced by Liang and Zeger (1986) as an extension of generalized linear models (GLM) to analyze discrete and correlated data. Its strength is that it models a known function of the marginal expectation of the dependent variable as a linear function of explanatory variables.

2 Purpose

The study will seek to examine an analysis of the academic performance of KNUST Level 400 Mathematics students admitted in the 2006 academic year in relation to certain corresponding variables influencing their performance. The general aim of the study is to use Generalized Estimating Equation (GEE) model analyses to compare the means of the identified variables affecting the performance of students with respect to the academic performance in their respective SWA scores. The study however has the following as it specific objectives:

1. to fit Generalized Estimating Equation (GEE) family of models under different working correlation assumptions to compare the means of students' Semester Weighted Average (SWA) in relation to their socio-demographic factors (such as gender, age), geographical location and graded level of former school attended.
2. to investigate whether these factors have effect on students' academic performance relating to their Semester Weighted Average (SWA) scores obtained in the University.

3 Methodology

A focus on the methodological review of Mathematical statistical tools that are relevant to the analyses of the various data gathered were used. Basically, the study seeks to use Generalized Estimating Equation (GEE) family of models, an extension of Generalized Linear Model (GLM) which takes into consideration Marginal Models for Longitudinal Data for the study.
The following statistical softwares such as, SPSS 16, Minitab version 14 and SAS version 9.1 were used. In addition, Semester results data of KNUST Mathematics students were useful for the analyses.

3.1 Data Source and Sample

The consecutive students' Semester Weighted Average (SWA) academic results from (2008-2011) of final year mathematics IV students at Kwame Nkrumah University of Science and Technology (KNUST) for each Semesters (i.e. seven Semesters) were obtained. The obtained SWA(s) scores were also tallied with the responses of sampled questionnaires for the students, requesting their gender, entry age, Grade level of former school attended as well as their geographical locations. The data variables were also coded in the Windows Microsoft excel 2007, SPSS version 17 and the SAS version 9.1 softwares were used for the analysis.
The entire population of MATH IV (2010/2011 academic year was obtained. A census sample size of 126 Mathematics students was sampled from Level 400 students in the Mathematics Department for the study. The researcher considered Level 400 students admitted into the University in 2006/2007 academic year based on their series of Semester examination relatively covering the whole requirements for their degree programme. They also have experience to share as far as various variables that affect their SWA are concern.

3.2 Variables

The study includes four manifest variables in the GEE analyses. These variables include gender, age, geographical location of students and former school attended. Gender was dummy coded. The geographical locations of students were categorized into four zonal belts in Ghana specifying their respective region of origin. These include the Northern Belt (comprising Northern , Upper East, Upper West Regions of Ghana) coded as L1, Middle/Central Belt (comprising Ashanti, Brong) was coded as L2, Eastern Belt (Eastern, Volta regions of Ghana) was coded as L3, and South/Coastal Belt (comprising Greater Accra, Central & Western regions) were also coded as L4.

Similarly, the graded schools were categorized into A, B and C, from Ghana Education Service (GES) specification. Grade A schools were coded as 1, grade B schools were coded as 2 and grade C schools were coded as 3

3.3 Data Processing

During the survey period, data were captured in person-level format and were reshaped into long format or person-period format so that it could be inputted to SAS PROC GENMOD for GEE analyses.. Students response to seven (7) questionnaires together with their SWA scores over the seven semesters covered in the university were relevant to the study. To determine whether there was an underlying structure, first the correlation matrix was examined to determine its appropriateness for the correlational analysis. The correlation matrix contained sufficient covariation for factoring. The data were then subjected to four GEE working correlation assumptions specified for the study.

4 Models Used For Anlaysis

Since students academic performance over the semesters (SWA) scores was correlated, the assumption under OLS regression was violated. As such these correlations needed to be taken into account in modeling; otherwise the standard errors of the estimates would be underestimated for the between-subject and overestimated for the within-subject effects. Generalized estimating equations (GEE) were introduced by Liang and Zeger (1986) as an extension of generalized linear models (GLM) to analyze discrete and correlated data. Its strength is that it models a known function of the marginal expectation of the dependent variable as a linear function of explanatory variables. The advantage of GEE is especially obvious when the number of observations is large in relative to number of waves within subjects as in the case of our dataset. Although there are other statistical procedures discussed in the literature which also take into account of correlated responses such as Weighted Least Squares (WLS), we will not be discussing these techniques as they are beyond the scope of this paper except to highlight that compared to WLS, GEE allows us to use continuous variables in the model while WLS does not. This flexibility makes GEE a better choice than WLS when modeling longitudinal data as WLS is restricted to categorical data in modeling. The following subsections further elaborates the GEE modeling.

4.1 Generalized Estimating Equations (Gee) Models

Let $Y_{ij};\ j = 1, \ldots, n_i,\ i = 1, \ldots, k$ K represent the j^{th} measurement on the i^{th} subject.

There are n_i measurements on subject i and $\sum_{i=1}^{k} n_i$ total measurements. Correlated data are modeled using the same link function and linear predictor setup (systematic component) as the independence case. The random component is described by the same variance functions as in the independence case, but the covariance structure of the correlated measurements must also be modeled. Let the vector of measurements on the i^{th} subject be $Y_i = \mu_i = [Y_{i1}, \ldots, Y_{ini}]'$ with corresponding vector of means $\mu_i = [\mu_{i1}, \ldots, \mu_{ini}]'$ and let V_i be an estimate of the covariance matrix of Y_{ij}. The Generalized Estimating Equation for estimating β is an extension of the independence estimating equation to correlated data and is given by

$$\sum_{i=1}^{k} \frac{\partial \mu_i}{\partial \beta} V_i^{-1}(Y_i - \mu_i(\beta)) = 0$$

A basic feature of GEE models is that the joint distribution of a subject's response vector y_i, does not need to be specified. Instead, it is only the marginal distribution of y_{ij} at each timepoint that needs to be specified. To clarify this further, suppose that there are two timepoints and suppose that we are dealing with a continuous normal outcome. GEE would only require us to assume that the distribution of y_{i1} and y_{i2} are two univariate normals, rather than assuming that y_{i1} and y_{i2} form a joint bivariate normal distribution. Thus, GEE avoids the need for multivariate distributions by only assuming a functional form for the marginal distribution at each timepoint.

A related feature of GEE models is that the (co)variance structure is treated as a nuisance. The focus is clearly on the regression of y on X. In this regard, GEE models yield consistent and asymptotically normal solutions for the regression coefficients $\beta(s)$, even with misspecification of the (co)variance structure of the longitudinal data. Since GEE models can be thought of as an extension of GLMs for correlated data, the GEE specifications involve those of GLM with one addition. So, first, the linear predictor is specified as

$$\eta_{ij} = x_{ij}^{'} \beta \qquad (1)$$

Where x_{ij} is the covariate vector for subject name *i* at time *j*. Then we consider a link function given as

$$g(\mu_{ij}) = \eta_{ij} \qquad (2)$$

is chosen. As in GLMs, common choices here are the identity, logit, and log link for continuous, binary, and count data, respectively. The variance is then described as a function of the mean, namely,

$$V(\mu_{ij}) = \varphi v(\mu_{ij}) \qquad (4)$$

Where, again, $v(\mu_{ij})$ is a known variance function and φ is a scale parameter that may be known or estimated.

4.2 The Gee Estimation (Working Correlations)

Defining A_i to be the $n_i \times n_i$ diagonal matrix with $V(\mu_{ij})$ as the j^{th} diagonal element, as indicated above, we define $R_i(a)$ to be the $n_i \times n_i$ "working" correlation matrix (of the n repeated measures) for the i^{th} subject (i.e. Y_i). Then, the working variance-covariance matrix for Y_i equals

$$V(a) = \varphi A_i^{\frac{1}{2}} R_i(a) A_i^{\frac{1}{2}}. \quad (5)$$

For the case of normally distributed outcomes with homogeneous variance across time, we get

$$V(a) = \varphi R_i(a) \quad (6)$$

For normal outcomes, Park (1993) extends this to heterogeneous variance across time by allowing the scale parameter φ_j to vary across time *(j* = 1, . . . , *n).*
The GEE estimator of β is the solution of

$$\sum_{i=1}^{N} D_i'[V(\hat{a})]^{-1}(y_i - \mu_i) = 0 \quad (7)$$

Where $\hat{a}$ is a consistent estimate of a and $D_i = \left(\frac{\partial \mu_i}{\partial \beta}\right)$ and hence, equation (6) becomes

$$\Sigma_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right) (V(\hat{a}))^{-1}[y_i - \mu_i] = 0 \quad (8)$$

This is an extension of the estimating equation for β in any GLM, which is given in (7). Thus, the GEE solution can be seen as a natural generalization of the GLM solution for correlated data. A**s** an example, in the normal case, for equation (7), that is

$$U(\beta) = \sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right)' (V(y_i))^{-1}[y_i - \mu_i] = 0$$

$$\mu_i = X_i \beta \quad (9)$$

$$D_i = X_i$$

$$V(\hat{\alpha}) = R_i(\hat{\alpha})$$

The solution for the parameter β (by making β a subject) results in;

$$\beta = \left[\sum_{i=1}^{N} X_i' [R_i(\hat{\alpha})]^{-1} X_i\right]^{-1} \left[\sum_{i=1}^{N} X_i' [R_i(\hat{\alpha})]^{-1} y_i\right] \qquad (10)$$

These are quasi-likelihood estimates since the equation depends on the mean and variance of *y*. Solving the GEE involves iterating between the quasi-likelihood solution for estimating β and a robust method for estimating α as a function of β. Basically, it involves the:

1. Given estimates of $R_i(\alpha)$ and φ, calculate estimates of β using IRLS.
2. Given estimates of β, obtain estimates of α and φ. For this, calculate Pearson (or Standardized) residuals

$$r_{ij} = \frac{(y_{ij} - \ddot{\mu}_{ij})}{\sqrt{[V(\hat{\alpha})_{jj}]}} \qquad (11)$$

and use these residuals to consistently estimate $\boldsymbol{a}$ and φ. Liang and Zeger 119861 present the estimators for several different working correlation structures. Upon convergence, in order to perform hypothesis tests and construct confidence intervals, it is of interest to obtain standard errors associated with the estimated regression coefficients. These standard errors are obtained as the square root of the diagonal elements of the matrix $V(\hat{\beta})$ .The GEE provides two versions of these:

1. **Naive or "model-based" estimator:**

This is the GEE equivalent of the inverse of the Fisher information matrix that is often used in generalized linear models as an estimator of the covariance estimate of the maximum likelihood estimator of $\hat{\beta}$.

$$V(\hat{\beta}) = \left[\sum_{i=1}^{N} D_i' \left(\hat{V}_i^{-1}\right) D_i\right]^{-1} \quad \textit{and for } D_i = X_i \textit{ becomes } V(\hat{\beta})$$

$$V(\hat{\beta}) = \left[\sum_{i=1}^{N} X_i' \hat{V}_i^{-1^{-1}} X_i\right]^{-1} \qquad (12)$$

It is a consistent estimator of the covariance matrix of $\hat{\beta}$ if the mean model and the working correlation matrix are correctly specified.

2. **Robust or "empirical" or sandwich estimator:**

The estimator

$$V(\hat{\beta}) = \sum_{i}^{N} M_0^{-1} M_1 M_0^{-1} \qquad (13)$$

is called the empirical, or robust, estimator of the covariance matrix of $\hat{\beta}$. It has the property of being a consistent estimator of the covariance matrix of $\hat{\beta}$ even if the working correlation matrix is misspecified, where

$$\mathrm{M}_0 = \left[\sum_{i=1}^{N} D_i'(\hat{V}_i^{-1})\, D_i\right]^{-1}$$

$$\mathrm{M}_1 = \sum_{i=1}^{N} D_i'\hat{V}_i^{-1}\,(y_i - \hat{\mu}_i)\,(y_i - \hat{\mu}_i)'\hat{V}_i^{-1}D_i$$

$$V(\hat{\beta}) = \left[(\sum_{i=1}^{N} D_i'(\hat{V}_i^{-1})\, D_i)^{-1}\right]^{-1} \times [\sum_{i=1}^{N} D_i'\hat{V}_i^{-1}\,(y_i - \hat{\mu}_i)\,(y_i - \hat{\mu}_i)'\hat{V}_i^{-1}D_i]$$

$$\times \left[(\sum_{i=1}^{N} D_i'(\hat{V}_i^{-1})\, D_i)^{-1}\right]^{-1}$$

Here, $\hat{V}_i$ denotes $\hat{V}_i(\alpha)$. We notice that if $\hat{V}_i = (y_i - \hat{\mu}_i)\,(y_i - \hat{\mu}_i)'$ then the two are equal. This occurs only if the true correlation structure is correctly modeled. Generally, we can deduce that, *the robust* or *"sandwich" estimator,* which is due to (Royal1, 1986), provides a consistent estimator of $V(\hat{\beta})$ even if the working correlation structure $R_i(a)$ is not the true correlation of yi

4.3 Specifying the Working Correlation Matrix

In GEE modeling, one has to specify the working correlation matrix in estimating the covariance of the parameter estimates. The specification of the working correlation matrix accounts for the form of within-subject correlation of responses on dependent variables. One of the aims of this paper is to find out whether using different working correlation matrices for estimation would affect the estimates and SEs (with respect to model-based and empirical-based) substantially.
Let **R**i(α) be an ni x ni working correlation matrix that is fully specified by the parameter α, the covariance matrix of Y is modeled as follows: Four types of working correlation were examined in this study in order to measure the relation-

ship between the student's SWA scores over time(seven semesters). They are briefly summarized below and examples given are presented in matrix form for the seven semesters.

Working correlation assumption	Correlation matrix
Independence GEE assumes that there is no correlation within the clusters of students' SWA scores and the model becomes equivalent to standard normal regression	$Corr(y_{ij}, y_{i,j+k}) = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} = I$
Exchangeable GEE working correlation specification allows for constant correlations between any two (2) measurements of the SWAs within a subject for all the time points (across the seven semesters).	$Corr(y_{ij}, y_{ii}) = \begin{cases} 1, & j = k \\ \rho, & j \neq k \end{cases}, \begin{pmatrix} 1 & \rho & \rho & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & \rho & \rho & 1 \end{pmatrix}$
Autoregressive GEE weights the correlation within SWA scores by their separated time over the seven semesters.	$Corr(y_{ij}, y_{i+k}) = \rho^k, \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 & \rho^6 \\ \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 \\ \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^6 & \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$
Unstructured GEE working correlation structure assumes different correlations between any two SWAs for every student. No constraints are placed on the correlations. Every element of the correlation matrix is estimated separately.	$Corr(y_{ij}, y_{i,k}) = \begin{cases} 1, & j = k \\ \rho_{jk}, & j \neq k \end{cases}, \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} & \rho_{15} & \rho_{16} & \rho_{17} \\ \rho_{21} & 1 & \rho_{23} & \rho_{24} & \rho_{25} & \rho_{26} & \rho_{27} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} & \rho_{35} & \rho_{36} & \rho_{37} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 & \rho_{45} & \rho_{46} & \rho_{47} \\ \rho_{51} & \rho_{52} & \rho_{53} & \rho_{54} & 1 & \rho_{56} & \rho_{57} \\ \rho_{61} & \rho_{62} & \rho_{63} & \rho_{64} & \rho_{65} & 1 & \rho_{67} \\ \rho_{71} & \rho_{72} & \rho_{73} & \rho_{74} & \rho_{75} & \rho_{76} & 1 \end{pmatrix}$

4.4 Generalized Wald Tests for Model Comparison

In order to interpret the group-related effects, we compare these models statistically to determine if the group by time interaction terms is jointly significant or not. Because GEE model parameters are estimated using quasi-likelihood procedures, there is no associated likelihood underlying the model. To compare the above GEE models, however, one can construct a multi-parameter Wald test to test the joint null hypothesis that a set of βs equal 0. For this, we define a q x p indicator matrix C of ones and zeros to select the parameters of interest for the multi-parameter test. Here, p equals the number of regressors in the full model (including the intercept) and q equals the number of parameters in the multi-parameter test (i.e**.,** the difference in regressors between the full and reduced models). The multi-parameter or generalized, Wald test then equals

$$x^2 = \hat{\beta}'C'(CV(\hat{\beta})C')'C\hat{\beta} \qquad (14)$$

which is distributed as x^2 with q degrees of freedom under the null hypothesis. The prime symbol $'$ indicates the transpose of the matrix or vector. Where C is a 1 x p vector selecting a single regression coefficient β. This will help test the hypothesis that:

H_0: $\beta_1 = \beta_2 = \beta_3 = \cdots = \beta_q$ (e.i. H_1: $\beta_q = \beta_p$), against the alternative that

H_1: $\beta_q \neq \beta_p$

Working Correlation Example – 4 Waves …..is a dispersion parameter, is the working correlation Where an diagonal matrix n n with diagonal matrix with as the jth diagonal element.

SAS Output

The macro %SelectGEE generated the three measures in the following SAS output.

The macro also produced the following summarized output (Table 1) which was exported from SAS output to EXCELwith minor amendments in wordings.

5 Results

Table 1 summarizes the computations for parameter estimates for our GEE working correlation assumptions (Independent, exchangeable, unstructured and AR (1) used for our analyses. The parameter estimates for the four models are approximately the same. A parameter estimates that is asterisked shows a statistical significance effect of its estimation in the model at 5% level of significance. The standard error estimates for each assumption are two: the model-based and the empirical based are given for each model. The standard error that is bracketed (.) represents model-based standard error. All computations are approximated to three decimal places. Gender 0 and 1 represent male and female students respectively. LOC represents geographical location of students; SCH represents a short-cut for type of schools students attended.

Table 1 Estimated Coefficients, Standard Errors and P-Values : GEE Models Parameter

	INDEPENDENT		EXCHANGEABLE		UNSTRUCTURED		AR(1)	
Parameter	**Estimate**	**Std Err**	**Estimate**	**Std Err**	**Estimate**	**Std Err**	**Estimate**	**Std Err**
INTERCEPT	59.421*	5.023*(4.311)*	59.97*	5.273*(5.273)*	59.420*	5.1824*(4.311)*	59.420*	4.312(*5.146*)
AGE	-0.034	0.203*(0.171)*	-0.034	0.215*(0.213)*	-0.034	0.203*(0.171)*	-0.034	0.213(*0.169*)
GENDER 0	-1.180*	0.913*(0.968)*	-1.118*	0.951*(0.958)*	-1.180*	0.913*(0.968)*	-1.180*	4.975(*4.050*)
GENDER 1	0.000	0.000*(0.000)*	0.000	0.000*(0.000)*	0.000	0.000(*0.000*)	0.000	5.146(*4.184*)
LOC 1	-3.601*	1.512*(1.289)*	-3.672*	1.338*(1.587)*	-3.604*	1.512(*1.289*)	-3.656*	1.320(*1.259*)
LOC 2	-0.525	0.962*(0.845)*	-0.528	0.866*(1.010)*	-0.525	0.962(*0.845*)	-0.525	0.852(*0.801*)
LOC 3	-0.572	1.049*(1.179)*	-0.572	1.161*(1.101)*	-0.571	1.049(*1.179*)	-0.572	1.150(*0.873*)
LOC 4	0.0000	0.000*(0.000)*	0.000	0.000*(0.000)*	0.000	0.000(*0.000*)	0.000	0.000(*0.000*)
SCH *A*	1.175	1.115*(1.073)*	1.176	1.142*(1.171)*	1.175	1.115(*1.073*)	1.176	1.124(*0.929*)
SCH *B*	0.283	1.172*(1.130)*	0.285	1.202(*1.231)*	0.283	1.172(*1.130*)	0.283	1.178(*0.976*)
SCH *C*	0.000	0.000*(0.000)*	0.000	0.000*(0.000)*	0.000	0.000(*0.000*)	0.000	0.000(*0.000*)

* Shows a parameter estimate that has significant effects at 5% level of significance.

(.) Shows model based standard error (std err).

5.1 GEE Main Effect Model

Table 1and 2 gives the analyses of GEE parameter estimation for the main effect based on our four assumptions (independent, autocorrelation, unstructured and exchangeable) with their respective model-based and empirical-based standard error estimates. The specification of the GEE independent assumes that there is no correlation within the students' SWA's and the model becomes equivalent to standard normal regression. The parameters are estimated within *seven* SWA semester scores of each student over time or seven semesters. Thus, the dependent variables are uncorrelated and are independent of each other across time (for all the seven semesters).

It could be inferred from table 1 that the parameter estimates for all the variables (age, gender, geographical locations and type of former school attended) are approximately the same for both empirical and model-based parameter estimates for all the assumptions. However, the standard errors for the robust and naïve cases are marginally different. This may indicate that the true correlation structure for the GEE is not correctly modeled using the independent, unstructured and autocorrelation model assumption.

The estimation of the model age parameter (-0.0542) was seen to be statistically insignificant at $\propto = 0.05$ significance level for all the assumptions. The parameter estimation for gender status of students is highly significance but has the highest standard errors values for both empirical and model-based estimators. In the

parameter estimate for location, only the difference between location 1 and location 4 (-3.8095) was significant with approximate standard error of 1.1935. The other contrasts for locations 2 and 3 were statistically insignificant and have estimation of standard errors different in the respective empirical and model based parameter estimates.

The contrast for type of former school attended A and B was also insignificant and have estimation of standard error different in the respective empirical and model based parameter estimates.

A critical observation of the standard errors for model-based and empirical-based estimators is marginally different and relatively smaller for all our assumptions except the exchangeable GEE model. The variations of these standard error estimates reduce the efficiency of considering the GEE unstructured working correlation assumption as not well fitted for the model. Only the contrasts within gender statuses 1 (female) of students as well as geographical location 1 and 4 are statistically significant in this model.

In the same way, information about the GEE Model (exchangeable) is displayed in table 1 and 2 for both empirical standard error and model-based standard error estimates respectively.

The parameter estimate for empirical and model based in the GEE exchangeable model are the same. The standard error estimate for empirical (robust or sandwich estimator) and model based are approximately the same. The parameter intercept is generally significant. The estimated standard error estimate for the robust and sandwich estimators of the model for all the parameter estimates are marginally equal. We notice that if

$$\hat{V}_i = (y_i - \hat{\mu}_i)\ (y_i - \hat{\mu}_i)$$

then the model-based and empirical-based standard errors estimates are equal. According to Royal1 [1986], this occurs only if the true correlation structure is correctly modeled. In this regard, comparing the analyses of the exchangeable GEE model with the other working correlation assumptions discussed above, we choose the exchangeable(compound symmetry) GEE model as the best fit for our analyses.

In our model including condition by linear time interactions with the independent parameters (see table 2), the parameter estimates were approximately the same. Only the intercept, location 1 and time interaction with type of formal school C were marginally significant at $\propto = 0.05$ for both the empirical and model based estimation. However, there is a negative trend for the parameter estimate for time interaction with some of the independent variables (males, location 1 and all the type of former schools students' attended Students' location 2 and 3) implying a decrease or diminishing trend of students SWA scores over some of the semesters (see fig 4.1 and 4.2).

Table2: GEE Model with Linear Time Interactions

	INDEPENDENT		EXCHANGEABLE		UNSTRUCTURED		AR(1)	
Parameter	**Estimat**	**Stand Err**	**Estimt**	**Std Err**	**Estimate**	**Std Err**	**Estimate**	**Std Err**
INTCPT	59.941*	5.155(*3.851*)	59.941*	5.154(*5.485*)	58.224	4.246(*5.182*)	60.518	5.100(*4.563*)
AGE	-0.054	0.215(*0.140*)	-0.054	0.215(*0.213*)	-0.053	0.172(*0.203*)	-0.054	0.213(*0.168*)
GEND 0	-1.369*	0.952(*0.631*)	-1.369*	0.951(*0.960*)	-1.369	1.009(*0.917*)	-1.368	0.942(*0.758*)
GEND 1	0.000	0.000(*0.000*)	0.000	0.000(*0.000*)	0.000	0.000(*0.000*)	0.000	0.000(0.000)
LOC 1	-4.180*	1.734(*2.207*)	-4.180*	1.733(*2.341*)	-4.183*	1.453(*2.053*)	-4.180*	1.683(*2.501*)
LOC 2	1.327	1.329(*1.480*)	1.327	1.328(*1.546*)	1.387	1.087(*1.348*)	1.325	1.257(*1.675*)
LOC 3	-0.071	1.529(*1.618*)	-0.071	1.529(*1.688*)	-0.071	1.445(*1.472*)	-0.071	1.523(*1.830*)
LOC 4	0.000	0.000(*0.000*)	0.000	0.000(*0.000*)	0.000	0.000(*0.000*)	0.000	0.000(*0.000*)
SCH *A*	-0.324	1.489(*1.700*)	-0.324	1.489(*1.780*)	-0.323	1.100(*1.554*)	-0.324	1.395(*1.924*)
SCH *B*	-0.856	1.517(*1.800*)	-0.856	1.517(*1.881*)	-0.865	1.166(*1.641*)	-0.856	1.431(*2.037*)
SCH *C*	0.000	0.000(*0.000*)	0.000	0.000(*0.000*)	0.000	0.000(*0.000*)	0.000	0.000(*0.000*)
TME*LOC 1	0.134	0.271(*0.397*)	0.134	0.271(*0.350*)	0.134	0.175(*0.342*)	0.134	0.251(*0.441*)
TME*LOC 2	-0.506	0.288(*0.365*)	-0.509*	0.287(*0.322*)	-0.509	0.206(*0.315*)	-0.509*	0.258(*0.406*)
TME*LOC 3	-0.129	0.341(*0.413*)	-0.128	0.341(*0.365*)	0.129	0.217(*0.356*)	-0.128	0.313(*0.460*)
TME*LOC 4	0.008	0.377(*0.416*)	0.008	0.376(*0.367*)	0.008	0.250(*0.358*)	-0.008	0.343(*0.462*)
TME* SCH *A*	0.360	0.307(*0.379*)	0.360	0.307(*0.334*)	0.361	0.204(*0.326*)	0.360	0.276(*0.421*)
TME* SCH *B*	0.303	0.346(*0.402*)	0.303	0.346(*0.355*)	0.302	0.229(*0.346*)	0.303	0.317(*0.447*)
TME* SCH C	0.000	0.000(*0.000*)	0.000	0.000(*0.000*)	0.000	0.000(*0.000*)	0.000	0.000(*0.000*)

5.2 GEE Model with Linear Time Interactions

From table 2, we model linear time interaction with the independent variables. The parameter estimate for both model base and empirical based were noted to be approximately the same with a significant intercept of 58.224[4.246(*5.182)*]. We note again a variation between the standard errors for the robust and naïve.

The model shows that, some of the linear trend parameter estimates are negative, (i.e. a decelerating negative trend is indicated). This may indicate a decrease in the average SWA scores within and between some of the semesters. Performance in SWA scores diminishes across time in the linear trend. Also, the effect of students' geographical location one is seen to be marginally significant, suggesting somewhat higher performance in their SWA scores across the seven semesters.

Apart from location one, all the other parameter estimates seem to have no statistical significance in the model estimation at 5% level of significance. This may imply that, the time interaction effects of the independent variables may not be necessarily contributing to the SWA scores of students in the KNUST Mathematics department.

In GEE AR (1) Model with Linear Time Interactions, the results of the analyses from table 2 is relatively similar to almost all the GEE family of models used for this study. In the same way, the interaction of location 1, 2 and 3 with location 4 are not statistically significant with variations in the standard error estimate for both model based and empirical based models.

Again, time interaction with locations and type of schools with formal school 3 record no significant level at $\propto = 0.05$ with variations in model-based and empirical standard error estimate for the independent, AR-1 and unstructured model. In linear time interactions with the exchangeable parameters as seen table 2, the parameter estimates were approximately the same.. However, there is a negative trend for the parameter estimate for time interaction with some of the independent variables such as gender, location 1 location 3 and time interaction with locations 2 and 3 were negative. Because some of the linear trend parameter estimates are negative, a decelerating negative trend is indicated. This may imply that, the time interaction effects of the independent variables may not be necessarily contributing to the SWA scores of students in the KNUST Mathematics department.

The standard errors for both model-based and empirical based in the independent GEE with Linear Time Interactions model are approximately the same for the model-based and empirical based reassuring our preference of considering its working correlation assumptions to the other

6 Conclusion

This study reaffirms the consistent estimate of GEE with the various working correlation matrix. Although the measures used in the study did not show the same results in the model selection process, they nevertheless provided useful guidelines and supported empirically that the specification of different working correlation pattern in the study did not differ much in their interpretations. The results of fitting the model parameter estimates were approximately identical, but the standard errors for the various GEE model varies within and across the parameters. This reaffirms the closeness of the measures for comparison of students' mean score in the KNUST Mathematics department. We note the closeness of the standard errors for both empirical and model based in the GEE model for Exchangeable confirming its suitability for the actual regression model.

In general, no significant effects were found for age and former school students attended. The study also revealed no gender difference in terms of academic performance based on SWA as confirmed by (Evans, 1999)

We noticed a statistically significance effect of students geographical locations in the model parameter estimation. The contrast effect of students from LOC1 (Northern, Upper East, Upper West) and LOC 2 (comprising Ashanti and Bono

Ahafo regions) with LOC 4(comprising Greater Accra, Western and Central regions) were highly significant. We conclude that, on the average, students from LOC4 score high SWA than the other two locations and hence they perform better. There was no significant effects between the difference of LOC1 and LOC4 with LOC3 (comprising Eastern and Volta regions).

From our findings, we recommend that Mathematics Education should be strengthened in the Northern, Upper East and Upper West regions of Ghana since the contrast effect of their average performance with other locations in Mathematics decreases over time.

References

[1] T. T. Kiang, K. Trivina, & D. Hogan, *Using GEE to Model Student's Satisfaction: A SAS ® Macro Approach,* Centre for Research in Pedagogy and Practice, Nanyang Technological University, Singapore (2009). Paper 251-2009 Practice, Nanyang

[2] J. W. Hardin, and J. M. Hilbe, *Generalized Estimating Equations*. Boca Raton, FL: Chapman and Hall/CRC Press, (2003).

[3] SAS Institute Inc., (*SAS/STAT User's Guide Volume 3, Version 9.1*. Cary, NC: SAS 2004).

Institute Inc.

[4] J. D. Singer and J. B. Willet, *Applied Longitudinal Data Analysis, Modeling Change and Event Occurrence*,(2003).

[5] M. Stokes, C. Davis, and G. Koch , *Categorical Data Analysis Using the SAS System*, 2nd Edition. SAS Institute, Cary, N. C. (2000).

[6] West African Examination Council, [WAEC], Chief Examiners' Report on 2008 students Academic Performance, WEAC Ghana.

[7] R.W.M. Wedderburn , *Quasi-likelihood functions, generalized linear models, and the gauss-newton method*. Biometrika, (1974). 61:439 {447, 1974}.

[8] , R. D. Wolfinger, Covariance structure selection in general mixed models. Communications in Statistics, Simulation and Computation. 22:1079-1 1 06,1993. York, 2000.

[9] S. L. Zeger & K. Y. Liang *Longitudinal data analysis using generalized linear models*. Biometrika, (1986) 73:13-22.

[10] S. L. Zeger, and , K. Y. Liang, *The analysis of discrete and continuous longitudinal data. Biometrics, 42*, (1986) 121-130.

Received: April 15, 2014