IOSR Journal of Research & Method in Education (IOSR-JRME)
e-ISSN: 2320–7388,p-ISSN: 2320–737X Volume 5, Issue 6 Ver. I (Nov. - Dec. 2015), PP 77-85
www.iosrjournals.org

A Longitudinal Study of Undergraduate Performance in Mathematics, an Application of Generalized Estimating Equation (GEE)

[1]Olukanye-David Oluwagbenga*, [1]Alo Damilola Olatubosun.
abestine@yahoo.com

[1]*Department of statistics, Federal University of Technology Akure, Nigeria*

Abstract: *Students' performance in mathematics has been an issue of great concern to most countries, especially the developing nations. So many programmes have been put in place to improve performances and to also encourage student to study the course in tertiary institution. In this study we investigate the relationship of semester, department of a student, age and load unit on marginal mathematics performance of undergraduate students. A marginal model was formulated using four working correlation structure where the exchangeable working correlation structure was selected as the best that models the dataset using quasi information criteria. The semester, age and load unit were found to be related to the marginal performance in mathematics*

Keywords: *longitudinal analysis, Generalized Estimating Equation, Repeated measure, marginal model, mathematics performance, marginal effect*

1. Introduction

In this study, we investigate the effect of some selected variables (semester, department, age and load unit) which is believed to affect students performance in mathematics using a repeated response. This research was as a result of degradation in students' performance in early university mathematics courses which affects their grade right from their inception in Federal University of Technology (FUT Akure) Nigeria. This study aim at answering the research question "Is students' marginal performance in mathematics related to semester, department, age and load unit?" understanding which of these variables that affect the marginal performance will assist the institution in policy that will enhance the students' marginal performance.

1.1 Background

Federal University of Technology Akure (FUTA) came into existence in September, 1981 but academic activities did not begin formally until November, 1982 with an enrolment of 149 students in three foundation schools, namely the School of Agriculture and Agricultural Technology (SAAT), the School of Pure and Applied Sciences (SPAS) now School of Science (SOS) and the School of Earth and Mineral Sciences (SEMS). Over the years, the number of schools has increased and the number of student increased to over 13,000. FUTA was adjudged the best University of Technology in Nigeria by the National University Commission (NUC) in 2004; produced the Nigerian best Researcher of the year in 2007; emerged the fifth best University in Nigeria in 2009 and was ranked among the best 50 Universities in Africa in 2011. FUTA became the centre of excellence in food security – a feat that attracted a grant of $700 million from World Bank.
In its quest to fulfil its vision and mission, it strives for academic excellence which birthed this paper.

The dataset used in this research is obtained both primarily and secondarily. The primary dataset is made of students' scores in mathematics in three consecutive semesters under study. The scores were monitored for three departments (biochemistry, physics and computer science) due to similarity in academic calendar and same mode of teach so as to reduce error that may arise as a result of disparity in activities which may lead to an invalid statistic. This collected variable formed a repeated measure over three time frames, which is also referred to as a longitudinal dataset.

The secondary part of the dataset comprises of students' bio-data information obtained from the school office. The information collected involves the age, gender of the student and the load unit for each semester was obtain from each departmental hand book.

Due to the importance of mathematics which serves as the bed rock of any developed country, many study has been channel towards looking into the cause of degradation in mathematics performance. Great deal of research has been carried out to determine how interest in mathematics can be improved from among young pupils, to do this several factors militating performance in mathematics at different level of studies have been investigated.

Olukanye and Ajiboye (2014) investigated the effect of some selected covariate over time on the performance of undergraduate student in mathematics. Their study was based on monitoring the performance of student from mathematics department for three consecutive semesters, where they discovered that load unit, time and the load unit over time are significant factors that affect mathematic performance.

Udousoro (2011) investigated the effect of gender and mathematics ability on academic performance of students in Chemistry. He studied the population of secondary school students and discovered that gender does not have any significant effect on students' performance in Chemistry. A similar research was conducted by Adeneye (2011), he considered the effect of gender on secondary school students' performance in Mathematics and discovered that gender has a significant effect on Mathematics performance.

This study would afford students the opportunity to be aware of certain existing variables that may have an impact on their performance in mathematics and hence improve themselves. Information gathered from results would inform students about their academic performance based on certain factors. Authorities in FUTA could also used the result to create a policy that will enhance the general performance in mathematics

The remaining sections in this paper are divided into four sections, section 2: Description of the Data and the Research Question; section 3: Methodology; section 4: Results and Discussion section 5: Conclusion.

2. Description Of Data And The Research Question

2.1 Data Description

The dataset used in this research is a real life dataset, it involves the study and follow up of students in three departments (Computer Science, Physics and Electronics and Biochemistry), from the duration of three semesters consecutively. These departments are selected because they belong to the same faculty (science) which brings about similarity in academic activities. The sample comprises of 92 students from computer science (35.5%), 85 students from physics (32.85) and 82 students from biochemistry (31.7%) which sum up to a total of 259 students. Three scores were obtained from each student which gives a longitudinal dataset of 777 observations measured in total.

The percentage composition of male and female in this study is as illustrated in table 2.1 below.

Table 2.1: Gender distribution

	BIOCHEMISTRY (%)	PHYSICS (%)	COMPUTER SCIENCE (%)	TOTAL
MALE	76 (89)	49 (60)	68 (74)	193
FEMALE	9 (11)	33 (40)	24 (26)	66
TOTAL	85 (100)	82 (100)	92 (100)	259

Also a brief summary of the explained variable across each department is displayed both in numeric representation and graphical representation to display hidden properties in the dataset. Table 2.2 below gives average performances of each department across the three semesters, their maximum score, minimum score and the standard deviation. From the table below, it can be seen that the students in computer science department performed better than the rest department in the first semester and third semester, where student in biochemistry department did well in the second semester. The least score was recorded in biochemistry department while the highest score was recorded in computer science in the first semester.

Table 2.2 Summary of Student Performance

SEMESTER	BIOCHEMISTRY		PHYSICS		COMPUTER SCIENCE	
	Mean (SD)	Min – Max	Mean (SD)	Min – Max	Mean (SD)	Min – Max
1	58.1 (13.2)	25 – 85	51.2 (14.2)	13 – 84	62.3 (14.2)	40 – 94
2	62 (12.1)	23 – 83	57.4 (15.1)	14 – 87	51 (12.1)	12 – 81
3	44.6 (15.6)	3 – 73	50.2 (13.7)	13 – 86	54.9 (11.7)	40 – 81

2.2 Variable used in this analysis

This sub-section describe the variables used in this study;

Demographic: the demographic variable used in this study is the gender. The variable identifies gender to which a particular student belongs. It is coded as (1) for male and (2) for female in this study.

Categorical variable: several categorical variables were also used, the semester is a categorical variable with 3 levels which is coded as (1) for the first semester, (2) for the second semester and (3) for the third semester. The department was also categorized into computer science, coded as (1), physics coded as (2) and biochemistry coded as (3).

Continuous variable: Two continuous variables were used which are the age and the load unit.

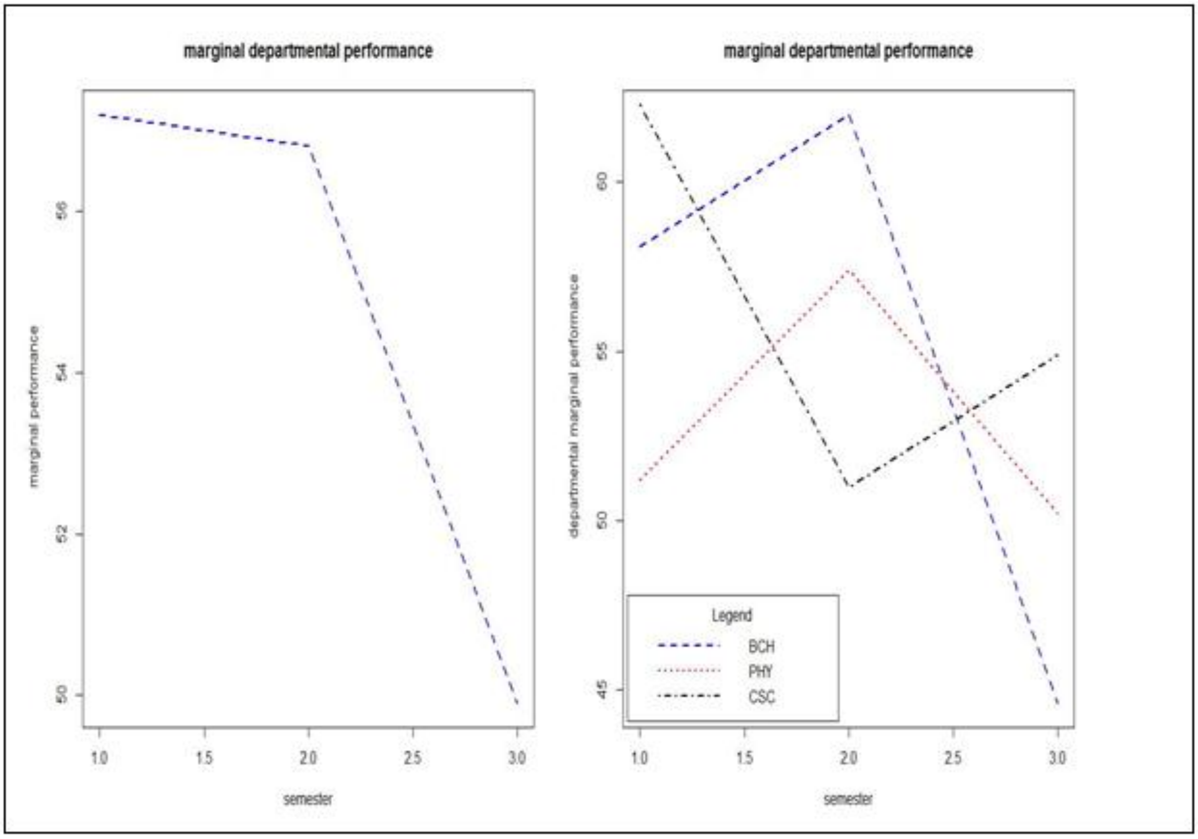
2.3 Marginal model

The specific objective of this research is the modelling of the marginal expectation of students' score as a function of these covariates: load unit, age, gender, department, semester and the interaction between load unit and semester. The marginal model which describes how the mean score relates to the covariate is given below;

$$E(Y_{it}) = \mu_{it} = \beta_0 + \beta_1 semester + \beta_2 dept_j + \beta_3 age + \beta_4 load\ unit + \beta_5 gender_j + \beta_6 (semester * load\ unit) \quad (2.1)$$

The marginal performance for each department and gender was shown in a graphical representation to examine the trend of performances across the time frame. Fig 2.1 reveals that the overall marginal performance decrease as the semester increases. The departmental marginal performance varies from department to department. The performances in biochemistry department tend to improve in the second semester which latter drop drastically in the third semester. The performance in computer science department is the reverse of that obtained in biochemistry department, while that of the physics department tend to be stationary between 50 and 60 marks.

Figure 2.1: Marginal Performances



3. Methodology

This research study is a descriptive study and the main design used was descriptive. However the mathematical methodology that was employed for the study was Generalized Estimating Equation (GEE) Model, i.e. a marginal model longitudinal (MML) data approach since the variable under study are multivariate with two or more dependents and independents variables.

3.1 Notation

Using the notation in Olukanye and Ajiboye (2014), we consider a study which involves N subjects, on which n observations are measured for each subject at T time points. Let $y_i = (\mathrm{y}_{i1}, \ldots, \mathrm{y}_{it})'$ denote the outcome measured for the ith subject associated with a vector of r $\times$ 1 covariates denoted by X_{iT}. Such data, known as longitudinal data, are found in different fields of life. This provides a means of studying the performance of any variable of interest over a time frame and also reduces within-subject error as a result of repeated measure. Such data exhibit a particular property (i.e. correlated) which needs to be accounted for in the course of analyses.

The variable of interest in this study is scores of students in mathematics, which is denoted by y_i, a (3 X 1) row vector. It comprises of students' score in three consecutive semesters in Introductory Mathematics I (MTS 101), Introductory Mathematics II (MTS 102) and Mathematical Methods I (MTS 201), which are all mathematics courses taught under the same conditions to the three departments.

The aim of this research was simplified into two specific research questions that can be answered, they are as follows;

1. Does each of the covariates considered affect the marginal performance in mathematics?
2. Which of the assumed working correlation best models the dataset?

Longitudinal dataset has a particular attribute which must be considered with care in order to choose an appropriate methodology in analysing. Considering the data structure in this study, the response variable measured on each student is collected for three consecutive time frames. This forms a cluster of responses on each student i.e. the responses are correlated. Hence the response measured in this study is a continuous correlated response and as such need to be account for. Also the interest in this study is the modelling of the expected response that is a marginal response as a function of some selected covariates. As a result of these two attributes in the dataset, generalised estimating equation (GEE) as discovered by Liang and Zeger (1986) is selected as a suitable methodology.

3.2 Generalised Estimating Equation

Generalized estimating equation was discovered by Liang and Zeger in 1986, in their quest to obtain a unified method suitable for the analysis of correlated responses of nay form (normal, binomial, count). GEE is an extension of Quasi-likelihood work of Wedderburn (1972). It is an extension of Generalized Linear model which accounts for dependency within responses. GEE can be used in two different analytical approach; subject-specific and population average (Liang and Zeger 1986). Population –Average also known as the marginal model was used in this study, because the aim of this research is the examination of the effect of some covariates on marginal response. This approach which measures the fixed effect of the covariates under study allows for specifying a predefined unique correlation structure often referred to as the working correlation structure which accounts for the dependency within subjects. The most often used working correlation structure and their construct are given below (Olukanye and Ajiboye 2014).

Table 3.1: Correlation Structures

Correlation type	Correlation formula	Working correlation structure
Independence	$Cor(Y_{ij}Y_{ik} = 0), j \neq k$	$R(\alpha) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}$
Exchangeable	$Cor(Y_{ij}Y_{ik} = \alpha), j \neq k$	$R(\alpha) = \begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & \alpha \end{pmatrix}$
AR(1)	$Cor(Y_{ij}Y_{ik} = \alpha^{\|j-k\|}), j \neq k$	$R(\alpha) = \begin{pmatrix} 1 & \alpha & \cdots & \alpha^{\|j-1\|} \\ \alpha & 1 & \cdots & \alpha^{\|j-2\|} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^{\|j-1\|} & \alpha^{\|j-2\|} & \ldots & 1 \end{pmatrix}$
Unstructured	$Cor(Y_{ij}Y_{ik} = \alpha_{jk}), j \neq k$	$R(\alpha) = \begin{pmatrix} 1 & \alpha_{12} & \cdots & \alpha_{1j} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{1j} & \alpha_{2j} & \cdots & 1 \end{pmatrix}$

As shown in the table above, the correlation structure R depends on α which can be estimated for exchangeable, ar1 and unstructured working correlation respectively from the following equations;

$$\hat{\alpha} = \varphi \sum_{i}^{n} \frac{1}{n_i (n_i - 1)} \sum_{j \neq k} R_{ij} R_{ik} \qquad (3.1)$$

$$\hat{\alpha} = \varphi \sum_{i}^{n} \frac{1}{(n_i - 1)} \sum_{j \leq n_{i-1}} R_{ij} R_{ij+1} \qquad (3.2)$$

$$\hat{\alpha}_{jk} = \varphi \tfrac{1}{n} \textstyle\sum_{i=1}^{n} R_{ij} R_{ik} \qquad (3.3)$$

Where the dispersion parameter estimate $\hat{\varphi}$ is given as;

$$\hat{\varphi} = \frac{1}{n-p} \sum_{i=1}^{n} \sum_{j=1}^{n_i} R_{ij}^2 \qquad (3.4)$$

In fitting a GEE model, several assumptions concerning the structure of the dataset accounts for the structure of the model construct. The response measured in this study is a continuous measure and hence is assumed to follow a normal distribution. This assumption allows choosing an identity link function which relates the covariates to the explained variable. The identity link function is given as;

$$g(\mu_i) = \mu_i = X_i\beta \qquad (3.5)$$

Where;

μ_i = average performance
X_i = is a matrix of predictors
β = regression coefficients

In order to achieve the aims and objectives of these studies, several parameter estimates need to be evaluated, especially the regression parameter β. Obtaining an estimate for β require solving equation (3.6) below called the score function or estimating equation, which is approached numerically (Liang and Zeger 1986)

$$U(\alpha,\beta) = \sum_{i=1}^{I} \frac{\partial \mu_i{}^T}{\partial\beta} V^{-1}(y_i - X_i\beta) \qquad (3.6)$$

Where;

$\frac{\partial \mu_i{}^T}{\partial\beta}$ = partial derivative of the μ_i w.r.t. each regression parameter in the model
$V = A^{-\frac{1}{2}}R(\alpha)A^{-\frac{1}{2}}$ (A variance-covariance matrix for a specified working correlation matrix)
$(y_i - X_i\beta)$ = the residual.

Unlike the conventional GLM which uses maximum likelihood estimation to obtain the parameters of the regression model, GEE uses estimating equation (equation 3.2). It uses an iterative procedure to obtain the regression parameter by assuming independency within cluster to obtain an initial regression parameter and then obtain an optimize regression parameter through iteration using any working correlation structure that best models the correlation within cluster. The most often used iterative equation is given below

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{m} + \left[\sum_{i=1}^{n}\left(\frac{\partial\hat{\mu}_i}{\partial\beta}\right)\hat{V}_i^{-1}\left(\frac{\partial\hat{\mu}_i}{\partial\beta}\right)\right]^{-1}\left\{\sum_{i=1}^{n}\left(\frac{\partial\hat{\mu}_i}{\partial\beta}\right)\hat{V}_i^{-1}(y_i - \hat{\mu}_i)\right\} \qquad (3.7)$$

Where;

$\hat{V}_i = V_i(\hat{\beta}^{(m)}, \hat{\alpha}(\hat{\beta}^{(m)}, \hat{\varphi}(\hat{\beta}^{(m)})))$ and $\left(\frac{\partial\hat{\mu}_i}{\partial\beta}\right)$ are also evaluated at $\hat{\beta}^{(m)}$. The $\hat{\beta}^{(m)}$ which serves as the initial value for the regression parameter is obtained from the Generalised Linear Model Method (GLM).

3.2.1 Iterative Process For GEE's

The procedure for the estimation of the regression parameter follows an iterative process as given below:

- Obtain the initial parameter assuming the response are uncorrelated (i.e. independent) using OLS and the dispersion parameter $\varphi = 1$
- Use the estimate β_{GLM} to calculate fitted values $\hat{\mu}_i = g^{-1}(X_i\beta)$.
- compute the Pearson residuals R_{ij} and obtain the estimates for φ, α and the working variance-covariance matrix V_i
- Using the current estimates $\hat{\alpha}, \hat{\varphi}$ and $\hat{\beta}$ in the Newton-Raphson iterative method to obtain a new improved regression parameter estimate

The iterative process is repeated until the regression parameter converges.

4. Results And Discussion

Having ascertained GEE to be a suitable method for analysis, the computation of all parameter estimates were done using R programming language. Assuming MCAR, geepack package was used in estimating the regression parameters and was also used to obtain some exploratory analyses both numeric and graphics.

For all four models fitted, there is no strong difference between independence and AR(1) correlation structure except the fact that age does not significantly affect the performance of students under independence correlation structure while it affects the performance under exchangeable correlation structure.

4.1 Assumptions

The following assumptions were used in fitting the model for this study;

1. **Link Function**
   Since the response (score) measured is a continuous variable which is assumed to follow a normal distribution, the identity link function is used.

i.e. $g(\mu_{ij}) = \mu_{ij}$

2. **Correlated Response**
   The response measured must be dependent, this is one of the basic assumptions under which GEE operates. The score collected for each student satisfies this assumption, since multiple scores (score 1, score 2, score 3) are collected for each student

These are the basic assumptions for the model in equation (2.1)

4.2 Results for Independence Correlation Structure

The following table contains the summary of the estimates obtained using R programming language under an assumed independence working correlation structure:

Table 4.1: Parameters Obtained from Independence Working Correlation GEE analysis

Effect	Estimate	Std. Error	Wald	Pr(>|W|)
Intercept	-5.0620	10.9218	0.21	0.643
Semester	10.6139	5.1552	4.24	0.040*
Physics department	2.9871	2.1524	1.93	0.165
Biochemistry department	0.5753	1.7206	0.11	0.738
Age	0.0676	0.0459	2.17	0.140
Load Unit	2.9543	0.4848	37.13	1.1e-09***
Gender (female)	-0.2782	1.5003	0.03	0.853
Semester * Load Unit	-0.6298	0.2507	6.31	0.012*

For the demographic variable considered, there was no significant effect for gender. As the semester increases, the performance of student increases significantly by approximately 11 marks. The load unit which is of paramount interest tends to increase the performance of student by approximately 3 marks which is also significant.

The load unit over time which is obtained as the interaction between semester and load unit also shows a significant negative effect on students' performance.

4.3 Results for Exchangeable Correlation Structure

The following table contains the summary of the estimates obtained using R programming language under the assumption that the correlation between the responses in a cluster is constant.

Table 4.2: Parameters Obtained from Exchangeable Working Correlation GEE analysis

Effect	Estimate	Standard err	Wald	Pr(>|W|)
Intercept	-4.6530	10.8408	0.18	0.6678
Semester	10.5776	5.1517	4.22	0.0400*
Physics department	2.9978	2.1526	1.94	0.1637
Biochemistry department	0.4056	1.7142	0.06	0.8129
Age	0.0564	0.0207	7.40	0.0065**
Load Unit	2.9471	0.4845	37.00	1.2e-09***
Gender (female)	-0.1755	1.4992	0.01	0.9068
Semester * Load Unit	-0.6287	0.2505	6.30	0.0121*

This result reveals that four factors out of all the predictors considered have significant effect on students' performance. With the estimates very similar to those obtained under independence assumption of the correlation structure. The estimated correlation matrix is given as:

$$\begin{pmatrix} 1 & 0.446 & 0.446 \\ 0.446 & 1 & 0.446 \\ 0.446 & 0.446 & 1 \end{pmatrix}$$

4.4 Results for Autoregressive Correlation Structure

The following table contains the summary of the estimates obtained using R programming language under the assumption that the observations obtain in close time frame are more correlated than those obtained in far time apart.

Table 4.3: Parameters Obtained from Autoregressive Working Correlation GEE analysis

Effect	Estimate	Standard err	Wald	Pr(>|W|)
Intercept	-12.2506	10.9288	1.26	0.2623
Semester	11.6269	5.3400	4.74	0.0295*
Physics department	3.0630	2.1699	1.99	0.1581
Biochemistry department	-1.6582	1.7644	0.88	0.3473
Age	0.0580	0.0224	6.74	0.0094**
Load Unit	3.3066	0.4934	44.91	2.1e-11***
Gender (female)	-0.1658	1.5464	0.01	0.9146

Semester * Load Unit	-0.6700	0.2601	6.64	0.0100**

The result obtained above also reveals that four of the variables considered have a significant marginal effect on the students' performance. As the student stays longer in the school system, the performance increases by approximately 12 marks, age also have a significant positive effect of about 0.1 marks as the age of the student increases. The load unit also have a positive significant increase of 3 marks as the load unit increases. The interactive effect of semester and load unit which measure the long run effect of load unit has a significant negative effect on the student performance.

The correlation structure estimate is given as;

$$\begin{pmatrix} 1 & 0.537 & 0.288 \\ 0.537 & 1 & 0.537 \\ 0.288 & 0.537 & 1 \end{pmatrix}$$

4.5 Results for Unstructured Correlation Structure

The following table contains the summary of the estimates obtained using R programming language under the assumption that the correlation within a cluster has no particular pattern. This is very close to the actual correlation.

Table 4.3: Parameters Obtained from Unstructured Working Correlation GEE analysis

Effect	Estimate	Standard err	Wald	Pr(>|W|)
Intercept	-4.4205	10.8366	0.17	0.6833
Semester	10.6532	5.1462	4.29	0.0384*
Physics department	3.0369	2.1484	2.00	0.1575
Biochemistry department	0.5887	1.7127	0.12	0.7311
Age	0.0556	0.0193	8.32	0.0039**
Load Unit	2.9319	0.4843	36.66	1.4e-09***
Gender (female)	-0.2018	1.4960	0.02	0.8927
Semester * Load Unit	-0.6303	0.2502	6.35	0.0118*

The result obtained above also reveals that four of the variables considered have a significant marginal effect on the students' performance. As the student stays longer in the school system, the performance increases by approximately 11 marks, age also have a significant positive effect of about 0.1 marks as the age of the student increases. The load unit also have a positive significant increase of 3 marks as the load unit increases. The interactive effect of semester and load unit which measure the long run effect of load unit has a significant negative effect on the student performance.

The correlation structure estimate is given as;

$$\begin{pmatrix} 1 & 0.452 & 0.487 \\ 0.452 & 1 & 0.400 \\ 0.487 & 0.400 & 1 \end{pmatrix}$$

4.6 Model Diagnostics

Since four different models are fitted using four assumed working correlation matrices, the need to select the best model that fit the data is necessary, even though the results obtained by these four models look similar. Quasi Information Criteria (QIC) which is a special kind of information criteria used to select model formed from quasi-likelihood. The rule of thumb selects the model with the least QIC value. The QIC value obtained for the four working correlation structure are (independence, autoregressive order 1, exchangeable and unstructured) are 150855.300, 151359.814, 150860.934 and 153272.179 consecutively. The smallest of all these is the independence working correlation structure, but due to the theoretical structure of the dataset which indicate presence of correlation, the least of the working correlation structure that accounts for correlation is adjudge the best, hence the model with exchangeable working correlation structure is considered the best model.

Table 9: Summary of GEE Models

Variables	GEE MODELS							
	Independent		Exchangeable		AR(1)		Unstructured	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
Intercept	-5.0620 (10.9218)	0.643	-4.6530 (10.8408)	0.678	-12.2506 (10.9288)	0.2623	-4.4205 (10.8366)	0.6833
Semester	10.6139 (5.1552)	0.040*	10.5776 (5.1517)	0.0400*	11.629 (5.3400)	0.0295*	10.6532 (5.1462)	0.0384*
Physics department	2.9871 (2.1524)	0.165	2.9978 (2.1526)	0.1637	3.0630 (2.1699)	0.1581	3.0369 (2.1484)	0.1575
Biochemistry department	0.5753 (1.7206)	0.738	0.4056 (1.7142)	0.8129	-1.6582 (1.7644)	0.3473	0.5887 (1.7127)	0.7311
Age	0.0676	0.140	0.0564	0.0065**	0.0580	0.0094**	0.0556	0.0039**

	(0.0459)		(0.0207)		(0.0224)		(0.0193)	
Load Unit	2.9543 (0.4848)	1.1e-09***	2.9471 (0.4845)	1.2e-09***	3.3066 (0.4934)	2.1e-11***	2.9319 (0.4843)	1.4e-09***
Gender (female)	-0.2782 (1.5003)	0.853	-0.1755 (1.4992)	0.9068	-0.1658 (1.5464)	0.9146	-0.2018 (1.4960)	0.8927
Semester *Load Unit	-0.6298 (0.2507)	0.012*	-0.6287 (0.2505)	0.0121*	-0.6700 (0.2601)	0.0100**	-0.6303 (0.2502)	0.0118*

Numbers in parentheses are robust standard errors

The inferences that could be made regarding the variable effects do not change substantially across the four models: examining GEE estimates from the different correlation structures reveals that those from the independence and exchangeable models are more identical compared to AR(1) and unstructured. However, four of the predictor variables significantly affects student's performance under exchangeable working correlation structure while only three variables significantly affects student performance under independent working correlation. All the variable have the same directional effect on the response variable but different not significant magnitude.

5. Conclusion

In this paper, we applied GEE to educational dataset using geepack package in R statistical programming language which assumes that any missing observation is missed completely at random (MCAR) to test the research hypothesis that students' performance in mathematics is related to semester, department, age and load unit. We found that load unit, semester, age and the interactive effect between semester and load unit affects the students' performance in mathematics. Based on the theoretical structure of the dataset used, the exchangeable working correlation matrix was adjudged the appropriate working correlation structure since it accounts for the dependency within scores obtained for each student. However student performance is not affected by gender and age category.

References

[1]. Adeneye, O, Adeleye,A (2011): "Is Gender a Factor in Mathematics Performance among Nigerian Senior Secondary Students with Varying School Organization and Location?" *International Journal of Mathematics Trends and Technology, vol. 2, Issue 3, 17 - 21*

[2]. Barnett, A.G., Koper, N., Dobson, A.J., Schimiegelow, F. and Manseau, M. (2010): "Using Information Criteria to Select the Correct Variance-Covariance Structure for Longitudinal Data in Ecology" *Methods in Ecology and Evolution*

[3]. Chaganty, N.R (1997): "An AlternativeApproach to the Analysis of Longitudinal Data via Generalized Estimating Equations" *Journal of statistical Planning and Inference*, 63, 39 – 54.

[4]. Chaganty, N.R and Joe, H (2004): "Efficiency of the Generalised Estimating Equations for Binary Response" *Journal of Royal Statistics* 66, 851 – 860

[5]. Cheong, Y. F., Fotiu, R. P. and Raudenbush, R. W. (2001): "Efficiency and Robustness of Alternative Estimator for Two and Three – level Models: The case of NAEP" *Journal of Educational and Behavioural Statistics*, 26, 411 – 429

[6]. Cox, D. R. (1972): "Regression Models and Life Tables (with discussion)" J*ournal of the Royal Statistical society* B,34, 187 – 220

[7]. Denis H. Y. Leung, You-Gan Wang and Min Zhu (2009): "Efficient parameter estimation in longitudinal data analysis using a hybrid GEE method". *Biostatistic*, 10, 3, pg 436 - 445.

[8]. Fitzmaurice, G.M., and Lipsits, S.R (2006): "Estimation in Regression Models for Longitudinal Binary Data with Outcome-Dependent Follow-Up" *Journal Biostatistics* 7,3, pp 469 - 485

[9]. Ghisletta, P and Spini, D (2004): "An Introduction to Generalized Estimating Equations and an Application to Access Selectivity Effects in a longitudinal Study on very Old Individuals" *journal of Educational and Behavioural Statistics* vol. 29, No.4, pp 421 – 437

[10]. Halekoh, U, Hojsagaard, S and Yan, J. (2006): " The R Package for Generalized Estimating Equations" *Journal of Statistical Software* vol. 15, No. 2 Pg 2 – 11

[11]. Hall, D.B and Severini, T.A (1998): " Extended Generalized Estimating Equations for Clustered Data " *Journal of the America Statistical Association* vol. 93, No. 444, Pg 1365 - 1375

[12]. Hay, J. L. and Pettitt, A.N (2001): "Bayesian Analysis of a Time Series of Counts with Covariates: An Application to the Control of an Infectious Disease" *Biostatistics* 2,4, pp 433 – 444

[13]. Hojnacki, M and Kimball, D.C (1998): "Organised Interests and the Decision of whom to Lobby In Congress" *American Political science Review* vol. 92, No. 4, 775 – 790

[14]. Huckfeldt, R, Sprague, J and Levine, J (2000): "The Dynamics of Collective Deliberation in the 1996 Election: Campaign Effects on Accessibility, Certainty and Accuracy" *American political Science Review* vol. 94, No. 3, Pg 641 - 651

[15]. James, A. H., Abdissa N., Micheal D. deb. Edwards and Janet E. Forrester (2002): "Statistical Analysis of correlated Data Using Generalized Estimating Equations: An Orientation". pp 364 - 375.

[16]. Jason R. (2003): "Scaled marginal model for multiple continuous outcome*" Biostatistics,* 4 ,3 , pp. 371 - 383

[17]. Johnson, P.E (2006): "Residuals and Analysis of fit: GLM #2 (version 2)". Pg 11 - 13

[18]. Joseph J.L and Alireza A. (2011): "An Overview of Longitudinal Data Analysis Methods for Neurological Research" *Journal of Dementia and Geriatric Cognitive Disorder Etra* vol. 1(1), 330 - 357

[19]. Kurland, B.F. and Heagerty, P.J (2005): "Directly Parameterized Regression Conditioning on Being alive: Analysis of Longitudinal Data Truncated by Deaths." *Journal of Biostatistics* 6,2, pp 241 – 258

[20]. Levitt, S.D (1996): "The Effect of Prison Population Size on Crime Rates: Evidence from Prison Overcrowding Litigation" *Quaterly Journal of Economics* vol. 111, Pg 319 - 352

[21]. Liang, K.Y and Zeger, S.L (2000): "Longitudinal Data Analysis of Continuous and Discrete Responses for Pre-Post Designs" *the Indian Journal of statistics* Vol. 62,B, pp 134-148

[22]. Liang, K.Y and Zeger, S.L (1986): “Longitudinal Data Analysis Using Generalized Linear Models.”*Biometrika*, 73, 13 - 32

[23]. Marco .G. and Matteo .B. (2007): “Quantile regression for longitudinal data analysis using the asymmetric laplace distribution” *Biostatistics* , 8 , 1 , pp. 140 - 154

[24]. Olukanye-David, O and Ajiboye, A.S (2014): “The Effects of a Set of Covariates on Mathematics Performance of Undergraduate Students: A Case Study of Mathematical Science FUTA” *The International Journal of Humanities and Social Studies*, Vol 2(12), 108 - 117

[25]. Oneal, J.R and Russett, B.M (1997): “The Classical Liberals were Right: Democracy, Interdependence and Conflict, 1950 – 1985” *International Studies Quaterly* vol. 41, No. 2, Pg 267 – 294

[26]. Pan, W. (2001), "Akaike's information criterion in generalized estimating equations," *Biometrics*, 57, 120-125

[27]. Peter M. (1983): “Quasi-likelihood Functions” *The Annals of statistics,* vol. 11, 1, 59 - 67

[28]. Richard, J.C and David, T. (2009): “Second-Order Estimating Equations for the Analysis of Clustered Current Status Data” *Journal of Biostatistics* 10,4, pp. 756 – 772

[29]. Robert .W. “Generalized Estimating Equations”, *Biostatistics* 411 ppt

[30]. Schildrout, J.S and Heagerty, P.J (2003): “Regression Analysis of Longitudinal Binary Data with Time-Dependent Environment Covariates: Bias and Efficiency” *Biostatistics* 6,4, pp 633 - 652

[31]. Scott P. N. (2009): “Using Generalized Estimating Equations (GEE) for Evaluating Research”ppt

[32]. Shults, J et. al., (2010): “Quasi-Least Squares with Mixed Linear Correlation Structures” *Journal of statistics and its interface*, 3, 223 – 233

[33]. Shults, J. and Chaganty, N.R (1998): “Analysis of Serially Correlated Data using Quasi-Least Squares” *Journal of Biometrics* 54: 1622 – 1630

[34]. Stram, D.O., Wei L. J. and ware J.H (1988): “Analysis of Repeated Ordered Categorical Outcomes with Possibly Missing Observations and Time-Dependent Covariates” *Journal of the American* pg 364 - 375.

[35]. Udousoro, U.J (2011):”The Effects of Gender and Mathematics Ability on Academic Performance of Students in Chemistry” *An International Multidisciplinary Journal, Ethopia vol. 5(4), No.21,Pg 201-213*

[36]. Wesley, K.T., Minge .X and Hellen R.W (2003): “Transformations of Covariates for Longitudinal Data” *Journal of Biostatistics* 4,3, pp. 353 – 364

[37]. Wickham, H. (2009): “ggplot2: Elegant Graphics for Data analysis” Springer New York

[38]. Yan, J (2002): “geepack: Yet Another Package for Generalized Estimating Equations” *R News* vol. 2, Pg 12 - 14

[39]. Yan, J. and Fine, J.P (2004): “Estimating Equations for Association Structures.” *Statistics in medicine.* Vol. 23, Pg 859 - 880

[40]. Yan, J., Aseltine, R. And Harel, O. (2011): “Comparing Regression Coefficients Between Nested Linear Models for Clustered Data with Generalized Estimating Equations” *Journal of Educational and Behavioural statistics* vol. 7, No. 2, Pg 101 - 121

[41]. Zeger, S.L and Liang, K-Y. (1986): “Longitudinal Data Analysis For Discrete and Continuous Outcomes”. *Biometrics* vol. 42, Pg 121-130

[42]. Zeger, S.L., Liang, K-Y and Paul S.A.(1986): “Models for Longitudinal Data: A Generalized Estimating Equation Approach” *Biometrics*, vol. 44 , Pg 1049 - 1060

[43]. Zorn, C.J.W. (2001): “Generalized Estimating Equation Models for Correlated Data: A Review with Applications” *American Journal of Political Science*, vol. 45, No. 2, Pg 470 – 490