THE INTERNATIONAL JOURNAL OF HUMANITIES & SOCIAL STUDIES

The Effects of a Set of Covariates on Mathematics Performance of Undergraduate Student: A Case Study of Department of Mathematical Sciences FUTA

Olukanye-David Oluwagbenga
Department of statistics, Federal University of Technology Akure, Ondo State, Nigeria
Ajiboye A. S.
Department of statistics, Federal University of Technology Akure, Ondo State, Nigeria

Abstract:
We employ the use of Generalized Estimating Equation in the longitudinal analysis to investigate the marginal effect of time, load unit, age category and gender on undergraduate performance in Mathematics. The analysis uses both descriptive and inferential statistics as tool to reveal hidden and confirmatory information.

The result reveals that time, load unit and the combined effect of time & load unit have a significant effect on the student performance in Mathematics. It also reveals that time and load unit has a positive effect on students' performance in Mathematics for all the models fitted under different working correlation structure.

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

1. Introduction

Nigerian educational system has undoubtedly faced many challenges which hinder most talented Nigerian student from attaining their full academic potentials. Education, the bed rock of every developed nation has been given more focus and attention in Europe and some of the African Nations. Hence many research works with different researchers' interests in different aspects have been carried out on students' performance in several fields of discipline.

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.

The interest of this research work is the performance of undergraduate students in Mathematics. We study the effect of gender, age category, load unit, time and the combined effect of time and load unit on the student performance in Mathematics.

1.1. Notation

Considering a study which involves n subjects, each measured at T time points, where $y_i = (y_{i1}, \ldots, y_{it})'$ denote the outcome measured for the ith subject associated with a vector of rx1 covariates denoted by X_{iT}. Such data, known as longitudinal data, are found in different fields of life. They exhibit a particular property (i.e. correlated) which needs to be accounted for in the course of analyses.

Our y_i = scores which is a row vector $(3 \mathrm{X} 1)$, 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).

2. Generalized Estimating Equation

Using the notations of Liang & Zeger (1986), given a row vector of response y_{it}repesenting the score of a student i at time t and the explanatory variables X_{it} representing the age category, gender, load unit, time and the interactive effect between time and load unit. The marginal response is defined as;

$$E(\mathrm{Y_i}) = \mu_i \qquad (1)$$

where the link function which relates the scores to the set of covariates is an identity. Identity link function is used since the response measured(scores) is a continuous variable which is assumed to have a Gaussian distribution. Using identity link function the general form of a marginal model is defined as;

$$\mu_i = g(\mathrm{X_i}\beta) \qquad (2)$$

where β is a k x 1 vector of parameters to be estimated.

Other forms of link functions available areLogit link function for Binomial response variable and Logarithm link function for Poisson response. Another important feature of Generalized Estimating Equation (GEE) that makes it the appropriate analytical choice in this study is the modelling of the within-subject correlation separately. Considering the construct of our dataset, three outcome variables (mathematics scores) for each student was obtained which formed a cluster i.e. a vector of observations. To account for the within-subject correlation, GEE uses four different correlation called “working correlation structure”.

Most used working correlation structures are Exchangeable, Autoregressive Order one (AR1) and Unstructured. These three correlation structures will be used to fit three models out of which the best model will be selected using two methods of model selection that are appropriate.

Obtaining the regression parameter in the marginal equation given in (2) above require solving of a score function which is defined as;

$$U(\beta\,,\alpha) = \sum_{i=1}^{n}\left(\frac{\partial\mu_i}{\partial\beta}\right)^T V_i^{-1}\left(y_i - \mu_i(\beta)\right) = 0 \qquad (3)$$

Where$V_i = A_i^{-\frac{1}{2}}R(\alpha)A_i^{-\frac{1}{2}}$ is the variance which incooperates the correlation model which is denoted by $R(\alpha)$. This value which is a matrix varies from structure to structure

GEE is an extension of Generalized Linear Model (GLM). This accounts for the within-cluster correlation. An important assumption of GLM is the independence of observations. This assumption is used to obtain the initial regression parameter β using ordinary least square (OLS) method which is then iterated over in Newton-Raphson iterative method given as;

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{m} + \left[\sum_{i=1}^{n}\left(\frac{\partial\hat{\mu}_i}{\partial\beta}\right)\hat{V}_i^{-1}\left(\frac{\partial\hat{\mu}_i}{\partial\beta}\right)\right]^{-1}\left\{\sum_{i=1}^{n}\left(\frac{\partial\hat{\mu}_i}{\partial\beta}\right)\hat{V}_i^{-1}(y_i - \hat{\mu}_i)\right\} \qquad (4)$$

where ;

$\hat{V}_i = V_i(\hat{\beta}^{(m)}, \widehat{\alpha}(\hat{\beta}^{(m)}, \widehat{\varphi}\,(\hat{\beta}^{(m)})))$ 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), i.e. the response from each observation are assumed to be uncorrelated (independent), then the Ordinary least square (OLS) is used to obtain $\hat{\beta}$ and then iterated to obtain a better regression parameter, i.e. ;

$$\hat{\beta}^{(0)} = (X^TX)^{-1}X^TY \qquad (5)$$

where $\hat{\beta}^{(0)} = [\hat{\beta}_1^{(0)} \vdots \hat{\beta}_2^{(0)} \vdots \cdots \vdots \hat{\beta}_p^{(0)}]$ and each $\hat{\beta}_i^{(0)}$ is a $(n \times 1)$ column vector i.e.

$$\hat{\beta}_1^{(0)} = \begin{bmatrix}\hat{\beta}_{10}^{(0)}\\ \hat{\beta}_{11}^{(0)}\\ \hat{\beta}_{12}^{(0)}\\ \vdots\\ \hat{\beta}_{1n}^{(0)}\end{bmatrix}\hat{\beta}_2^{(0)} = \begin{bmatrix}\hat{\beta}_{20}^{(0)}\\ \hat{\beta}_{21}^{(0)}\\ \hat{\beta}_{22}^{(0)}\\ \vdots\\ \hat{\beta}_{2n}^{(0)}\end{bmatrix}\hat{\beta}_3^{(0)} = \begin{bmatrix}\hat{\beta}_{30}^{(0)}\\ \hat{\beta}_{31}^{(0)}\\ \hat{\beta}_{32}^{(0)}\\ \vdots\\ \hat{\beta}_{3n}^{(0)}\end{bmatrix}\cdots\quad \hat{\beta}_p^{(0)} = \begin{bmatrix}\hat{\beta}_{p0}^{(0)}\\ \hat{\beta}_{p1}^{(0)}\\ \hat{\beta}_{p2}^{(0)}\\ \vdots\\ \hat{\beta}_{pn}^{(0)}\end{bmatrix}$$

Equation (5) can be re-written for each $\hat{\beta}_i^{(0)}$ as follows;

$$\begin{gathered}\hat{\beta}_1^{(0)} = (X^TX)^{-1}X^TY_1\\ \hat{\beta}_2^{(0)} = (X^TX)^{-1}X^TY_2\\ \vdots\\ \hat{\beta}_p^{(0)} = (X^TX)^{-1}X^TY_p\end{gathered}$$

Having obtained the initial β and looking at equation (4), there is a need to obtain the variance-covariance matrix which is a function of a working correlation $R(\alpha)$as earlier stated;this depends wholly on the parameterα . The estimate of this parameter is relative to the choice of the weighing scheme used to model the within-subject correlation as early stated. Hence, the consistent estimate for α under exchangeable, Autoregression of order one (AR(1)) and unstructured are given thus;

$$\hat{\alpha} = \varphi\sum_{i}^{n}\frac{1}{n_i(n_i-1)}\sum_{j\neq k}R_{ij}R_{ik} \qquad (6)$$

$$\hat{\alpha} = \varphi\sum_{i}^{n}\frac{1}{(n_i-1)}\sum_{j\leq n_{i-1}}R_{ij}R_{ij+1} \qquad (7)$$

$\hat{\alpha}_{jk} = \varphi\frac{1}{n}\sum_{i=1}^{n}R_{ij}R_{ik}$ (8)

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

$$\hat{\varphi} = \frac{1}{n-p}\sum_{i=1}^{n}\sum_{j=1}^{n_i}R_{ij}^2 \qquad (9)$$

2.1. Working Correlation Structure

The various correlation structure matrices are given below;

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

Table 1: working Correlation Structure

2.2. 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, at the convergence point, the best numerical regression parameter will be obtained. GEE uses different working correlation matrix to get different model, this calls for a way of selecting the model that best fits the data. Several methods are available to determine model goodness of fit, one of the most used methods is an equivalent Akaike Information Criteria known as Quasi Information Criteria.

3. Materials and Methods

3.1. Data Collection

Federal University of Technology Akure, Nigeria is one of the leading Universities in Nigeria, being the best Technology University. It is known for her academic excellence and cult-free activities. The University is situated in the capital city of Ondo State, established in the year 1981, commenced administrative activities in the 1982. At the moment, FUTA has six Schools (School of Sciences, School of Earth and Mineral Sciences, School of Environmental Technology, School of Engineering Technology, School of Agricultural and Agricultural Technology and School of Management Technology) with over thirty departments and about 10,000 students (8,000 Undergraduate and 2,000 postgraduate).

In this study, the School of Science is considered as the population of choice due to the similarity in activities and courses offered (mathematics) for the duration under study (4 months, 9 months and 16 months). Our sample is made up of 80 students from Department of Mathematical Sciences, offering the following Mathematics courses;

- Introductory Mathematics I (MTS101)
- Introductory Mathematics II (MTS102)
- Mathematical Methods MTS201

The scores for these students were obtained for three consecutive semesters which form our time frame (4 months, 9 months 16 months). The students under study offered these three courses at the same time in the same semester thereby giving a longitudinal data. The same means of teaching were used for all students and the same examination were administered to all the students under the same examination conditions.

3.2. Variables Used in the Analyses

This sub-section describe the various variables used in this research work

- Demographic: The demographic variable used in this analysis is gender. Gender indicates whether the student identifies as a male which is coded as (1) or as a female which is coded as (2)

- Categorical variable: This indicates the age category to which the student belong as at the time of admission as indicated in the student bio-data obtained from the school of science. The age category is divided into two sub-groups: less than or equal nineteen (≤ 19) which is classified as teen and coded as (1), while greater than or equal twenty (≥ 20) is classified as adult and coded as (2)
- Load Unit Variable: This indicates the load unit done in a particular semester. The load unit is also divided into two sub-groups: less than or equal nineteen (≤ 19) is classified as minimum and coded as (1), while the load unit greater than or equals twenty (≥ 20) is classified as maximum and coded as (2)
- Score Variable (y): This is a continuous variable which represent the score(s) of a student in a particular semester

A sample of 80 students was obtained from Mathematical Sciences Department (MTS), alongside their corresponding scores in these semesters and other information as shown in the appendix

3.3. Data Analysis

We used both exploratory and confirmatory statistical tool to analyze our dataset. Using some exploratory tools (both numerical and graphical) in a statistical programming language (R), we reveal some hidden information which will be discussed in this section.

Time	Teen		Adult	
	Mean (SD)	**Median (min - max)**	**Mean (SD)**	**Median (min - max)**
4 months	57.68 (9.96)	57.00 (40 - 81)	51.12 (14.51)	50.00 (16 - 78)
9 months	61.91 (11.01)	61.50 (40 - 78)	56.91 (16.85)	58.00 (07 - 85)
16 months	49.06 (12.50)	47.50 (19 - 72)	50.48 (15.74)	52.80 (11 - 78)

Table 2: summary statistics

The first numerical information obtained from our dataset is the summary statistics. This shows that the average performance of student who belong to the teen age category in Mathematics in the second semester improves (61.91) compare to that of the first semester (57.8), but drop in the third semester (57.8). Also, for the adult category, we discovered that the performance for this group of people lie within the range of 50 – 58. Though their performances follows a similar trend with that of teen group, but the third semester performance for the adult is better-off than the teen group.

This information was also presented in a line graph for better understanding of the performance trend for each group.

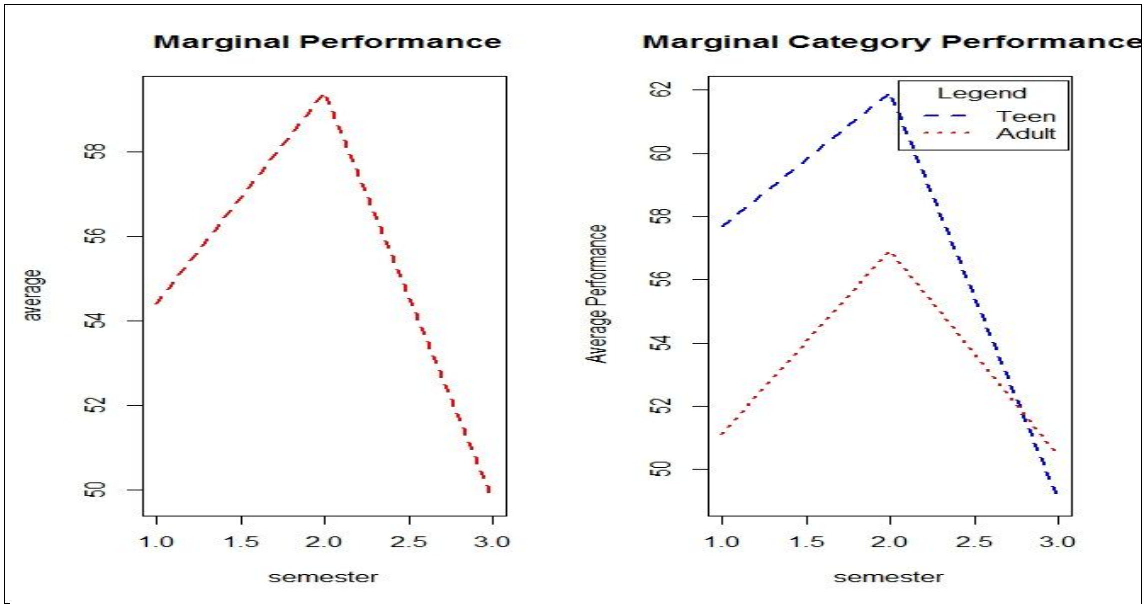

Figure 1: Marginal Trend Performance Plot

A box plot also was plotted for each semester, to visualize the performance of each group category as shown below;

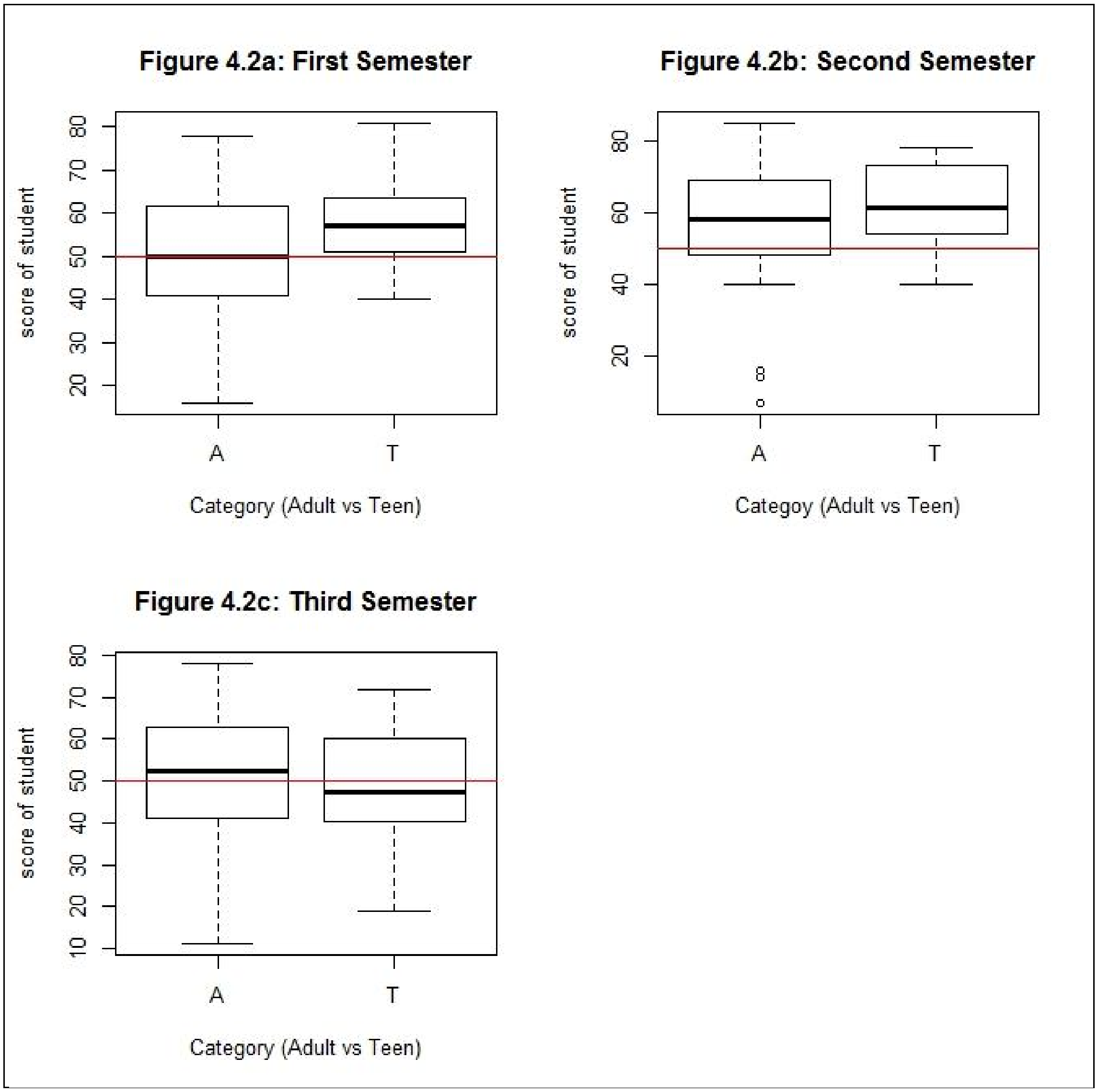

Figure 2: Box-Plot for each semester score

3.4. Model

We assume that our response (score) $Y_i = (Y_{i1}, \dots, Y_{in_i})^T$ follows Gaussian distribution since it is a continuous measure. The associated covariates $X_{ij} = (X_{ij1}, \dots, X_{ijp})^T$(Category, load unit, gender, time) are collected on student$i$, for $i = 1, \dots, 84$. The expected value and variance of measurement Y_{ij} can be expressed using generalized linear model (GLM):

Hence the model fitted is given as follows;

$$E(Y_{it}) = \mu_{it} = \beta_0 + Cat\beta_1 + Gender\beta_2 + Load\beta_3 + Time\beta_4 + (Time * load)\beta_5 \quad (10)$$

where;

Cat is a categorical explanatory variable with two levels (Adult (2), Teenager (1))

$Gender$ is a categorical explanatory variable with two levels (male (1), female (2))

$Load$ is a categorical explanatory variable (≤ 19 minimum (1)) (≥ 20 maximum (2))

Time is coded as a categorical explanatory variable with 3 levels (4month, 9months, 16months)

Load*Time is the effect of the load unit over time

$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are the parameters to be estimated

3.5. Assumptions

The following assumptions were used in fitting the model for this study as specified in equation (10) above

3.5.1. 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. $g(\mu_{ij}) = \mu_{ij}$

3.5.2. 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 (10). All the parameter was computed using statistical programming language (R), and the code is available in the appendix section for interested researcher.

4. Results

Analysis was completed in R with geepack package. The Four (4) working correlation matrices are used to fit four different models. The results are given in the tables below

	Estimate	Standard err.	Wald	Pr(>|W|)
Intercept	36.976	6.061	37.22	1.1 e-09
Category	-2.725	2.186	1.55	0.21
Time	1.927	0.390	24.38	7.9 e-07
Gender	1.232	2.872	0.18	0.67
Lu	18.265	3.666	24.83	6.3 e-07
Time*Lu	-1.616	0.314	26.57	2.5 e-07

Table 3: Independent working correlation matrix result

Scale parameter φ = 197.4

Independent Working Correlation matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Table 3 is the result generated under the assumption that the observation (scores) in each cluster (student) are independent. It shows that the following predictors; time, load unit and the interaction between load unit and time have a significant effect on the performance of student. The category coefficient (-2.725) indicates that Teen performs better than Adult (since they are categorical dummy variable, the estimates and the product of the dummy variable assigned to Adult will give a higher negative value than that assigned to Teen).

The time coefficient also indicates that as the student increase their stay on campus, their performance will increase semester increases by 1.927.

The coefficient for gender shows that female student performs better than male, though not statistically significant.

The coefficient for load unit indicates that a student with load unit greater than or equal to 20 will have a higher score that a student with load unit less than or equal 19.

The coefficient for the interaction between time and load unit shows that the performance of a student decreases with a coefficient of (-1.616) as the load unit increases over time

	Estimate	Standard err.	Wald	Pr(>|W|)
Intercept	36.704	6.0710	36.5519	1.487 e-09
Category	-2.437	2.1535	1.2804	0.2578
Time	1.686	0.3809	19.5969	9.562 e-06
Gender	1.768	2.8219	0.3925	0.5310
Lu	17.855	3.8001	22.0774	2.619 e-06
Time*Lu	-1.493	0.3293	20.5412	5.836 e-06

Table 4: Autoregressive order one (ar1) Working correlation matrix result

Scale parameter φ = 197.4

Autoregressive order 1 (ar1) Working Correlation matrix

$$\begin{pmatrix} 1 & 0.3308 & 0.1094 \\ 0.3308 & 1 & 0.3308 \\ 0.3308 & 0.1094 & 1 \end{pmatrix}$$

Table 5 is the result generated under the assumption that the degree of correlation within the cluster (student) decreases as the time between the observation increases. It shows that the following predictors: time load unit and the interaction between load unit and time have a significant effect on the performance of students. The category coefficient (-2.437) indicates that Teen performs better than Adult (since they are categorical dummy variable, the product of the dummy variable assigned to Adult and the estimate will give a higher negative value than that assigned to Teen).

The time coefficient also indicates that as the semester increases, the performance of the student increases by 1.686, i.e. the performance of each student in mathematics will improve over time.

The coefficient for gender shows that female student performs better than male, though not significant.

The coefficient for load unit indicates that a student with load unit greater than or equal to 20 will have a higher score than a student with load unit less than or equal 19.

The coefficient for the interaction between time and load unit shows that the performance of a student over time as the load unit increases decreases with a coefficient of (-1.493)

	Estimate	Standard err.	Wald	Pr(>|W|)
Intercept	36.636	6.182	35.12	3.1 e-09
Category	-2.727	2.176	1.57	0.21
Time	2.024	0.352	33.10	8.8 e-09
Gender	1.213	2.870	0.18	0.67
Lu	18.479	3.758	24.18	8.8 e-07
Time*Lu	-1.667	0.320	27.16	1.9 e-07

Table 5: Exchangeable Working correlation matrix result

Scale parameter φ = 197

Exchangeable Working Correlation matrix

$$\begin{pmatrix} 1 & 0.26 & 0.26 \\ 0.26 & 1 & 0.26 \\ 0.26 & 0.26 & 1 \end{pmatrix}$$

Table 6 is the result generated under the assumption that the degree of correlation within a cluster (student) is constant. compound symmetric (exchangeable). It shows that the following predictors: time, load unit and the interaction between load unit and time have a significant effect on the performance of student. The category coefficient (-2.727, which is very close to the value obtained under independent working correlation matrix) indicates that Teen performs better than Adult (since they are categorical dummy variable, the product of the dummy variable assigned to Adult and the estimate will give a higher negative value than that assigned to Teen).

The time coefficient also indicates that as the semester increases, the performance of the student increases by 2.024 that the performance of each student in mathematics will improve over time.

The coefficient for gender shows that female student performs better than male, which is also not statistically significant.

The coefficient for load unit indicates that a student with load unit greater than or equal to 20 will have a higher score that a student with load unit less than or equal 19.

The coefficient for the interaction between time and load unit shows that the performance of a student decreases with a coefficient of (-1.667) as the load unit increases over time

	Estimate	Standard err.	Wald	Pr(>|W|)
Intercept	35.605	5.947	35.85	2.1 e-09
Category	-1.873	2.182	0.74	0.39
Time	2.003	0.368	29.63	5.2 e-08
Gender	1.239	2.758	0.20	0.65
Lu	17.651	3.741	22.26	2.4 e-06
Time*Lu	-1.605	0.318	25.47	4.5 e-07

Table 6: Unstructured Working correlation matrix result

Scale parameter φ = 197.4

Unstructured Working Correlation Matrix

$$\begin{pmatrix} 1 & 0.5768 & 0.1316 \\ 0.5768 & 1 & 0.0976 \\ 0.1316 & 0.0976 & 1 \end{pmatrix}$$

Table 7 is the result generated under the assumption that the degree of correlation within a cluster is unique. It shows that the following predictors; time, load unit and the interaction between load unit and time have a significant effect on the performance of the students. The category coefficient (-2.727, which is very close to the value obtained under independent working correlation matrix) indicates that Teen performs better than Adult (since they are categorical dummy variable, the product of the dummy variable assigned to Adult and the estimate will give a higher negative value than that assigned to Teen).

The time coefficient also indicates that as the semester increases, the performance of the student increases by 2.024 that the performance of each student in mathematics will improve over time.

The coefficient for gender shows that female student performs better than male, which is not statistically significant.

The coefficient for load unit indicates that a student with load unit greater than or equal to 20 will have a higher score than a student with load unit less than or equal 19.

The coefficient for the interaction between time and load unit shows that the performance of a student decreases with a coefficient of (-1.667) as the load unit increases over time

4.1. 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. Three methods of selection will be considered which are grouped into two categories;

- Residual Analyses
- Information criteria

4.1.1. Residual Analyses

Residuals are frequently used to evaluate the validity of the assumptions of statistical models and may also be employed as tools for model selection (Nobre and da Motta Singer, 2007). The Marginal residual given by $\hat{\varepsilon} = y - X\hat{\beta}$ will form the basis on which the two residual analyses that will be employed in selecting the best model will be formed.

Under this category, two methods of model selection that are used are;

- Pearson Deviance which is given as $\sum_{i=1}^{n}(y - \hat{\mu})^2$
- Pearson chi-square statistics which is given as $\chi^2 = \frac{\sum_{i=1}^{n}(y-\hat{\mu})^2}{\phi V(\hat{\mu})}$

where;

$y - \hat{\mu}$is the Pearson residual

ϕ is the dispersion parameter

$V(\hat{\mu})$is the variance in our case it is 1 (from table 1.2)

4.1.2. Information Criteria

The information criteria used in this work is the extension of the Akaike Information criteria to quasi-likelihood model called the Quasi-Likelihood Information Criteria (QAIC or QIC). The results are given below

CORRELATION STRUCTURE	DEVIANCE	PEARSON CHI-SQUARE	QIC
Independence	46972	237.95	47040.2
AR(1)	46989	238.04	47052.6
Exchangeable	46973	237.96	47040.2
Unstructured	47043	238.31	47095.6

Table 7: Model diagnostic result

Variables	GEE MODELS			
	Independent	Exchangeable	AR(1)	Unstructured
Intercept	36.976 (6.061)	36.636 (6.182)	36.704 (6.071)	35.605 (5.947)
Category	-2.725 (2.186)	-2.727 (2.176)	-2.154 (2.154)	-1.873 (2.182)
Time	1.927 (0.390)	2.024 (0.352)	1.686 (0.381)	2.003 (0.368)
Gender	1.232 (2.872)	1.213 (2.870)	1.768 (2.822)	1.239 (2.758)
Lu	18.265 (3.666)	18.479 (3.758)	17.855 (3.800)	17.651 (3.741)
Time*Lu	-1.616 (0.314)	-1.667 (0.320)	-1.493 (0.329)	-1.605 (0.318)

Table 8: Summary of GEE Models

Cell entries are parameter estimates; numbers in parentheses are robust standard errors

The inferences one would make 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. This can be attributed to the estimated value of the correlation parameter in the exchangeable model which is relatively small (0.26), indicating only a low level of correlation among the score observed

5. Conclusion

Having assumed that the data that we used for analysis is MCAR, we found that load unit, time frame and the interaction of both are predictors. Based on our result, we discover that, Independence working correlation matrix may be the appropriate working correlation structured for repeated data measured show which a very small correlation. Student performance is not affected by gender and age category. However the load unit, time frame and the interaction of the two have a positive effect on the students' score.

6. 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)” Journal 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. Martinez, W.L., Martinez, A.R. and Solka, J.L: “Exploratory Data Analysis with Matlab”Second Edition pg 21.
25. McCullagh, P and Nelder, J. (1989): “Generalized Linear Models”, Second Edition. Chapman and Hall/CRC pg. 325
26. Nicholas H. (2001): “Fitting Generalized Estimating Equation (GEE) regression models in stata”.
27. Nobre, J.S.S and Motta Singer, J (2007): “Residual Analysis for Linear Mixed Models” Biometrical Journal, vol. 49, No. 6, Pg 863 – 875
28. 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
29. Pan, W. (2001), "Akaike's information criterion in generalized estimating equations," Biometrics, 57, 120-125
30. Paolo, G and Dario, S (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, Pg 421
31. Peter M. (1983): “Quasi-likelihood Functions” The Annals of statistics, vol. 11, 1, 59 - 67
32. 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
33. Robert .W. “Generalized Estimating Equations”, Biostatistics 411 ppt

34. 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
35. Scott P. N. (2009): "Using Generalized Estimating Equations (GEE) for Evaluating Research"ppt
36. Shults, J et. al., (2010): "Quasi-Least Squares with Mixed Linear Correlation Structures" Journal of statistics and its interface, 3, 223 – 233
37. Shults, J. and Chaganty, N.R (1998): "Analysis of Serially Correlated Data using Quasi-Least Squares" Journal of Biometrics 54: 1622 – 1630
38. Sojan G. (2009) "Advanced Longitudinal data analysis", John Hopkins University, Spring ppt
39. 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.
40. 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
41. Wesley, K.T., Minge .X and Hellen R.W (2003): "Transformations of Covariates for Longitudinal Data" Journal of Biostatistics 4,3, pp. 353 – 364
42. Wickham, H. (2009): "ggplot2: Elegant Graphics for Data analysis" Springer New York
43. www.onlinecourses.science.psu.edu/sta504/node/181
44. www.unc.edu/courses/2010spring/ecol/562/001\\/docs/lectures/lecture14.htm
45. Xie, J. And Shults, J. (2009): "Implementation of Quasi-Least Squares with the R Package qlspack UPenn" Biostatistics working Papers 32 http://biostats.bepress.com/upennbiostat/papers/art32
46. Yan, J (2002): "geepack: Yet Another Package for Generalized Estimating Equations" R News vol. 2, Pg 12 - 14
47. Yan, J. and Fine, J.P (2004): "Estimating Equations for Association Structures." Statistics in medicine. Vol. 23, Pg 859 - 880
48. 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
49. Zeger, S.L and Liang, K-Y. (1986): "Longitudinal Data Analysis For Discrete and Continuous Outcomes". Biometrics vol. 42, Pg 121-130
50. 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
51. 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