APPLICATION OF GENERALIZED ESTIMATING EQUATION (GEE) MODEL ON STUDENTS ACADEMIC PERFORMANCE: A CASE STUDY OF FINAL YEAR MATH FOUR STUDENTS AT KNUST

By

Owusu-Darko Isaac

A Thesis submitted to the Department of Mathematics, Kwame Nkrumah University of Science and Technology in partial fulfillment for the degree of

MASTER OF PHILOSOPHY, MATHEMATICS
Faculty of Physical Science.

JUNE, 2011

DECLARATION

I hereby declare that this thesis is my own work towards the Mphil and that, to the best of my knowledge, it contains no material previously published by another person nor material which has been accepted for the award of any other degree of the University or elsewhere, except where due acknowledgement has been made in the text.

(Student's name & ID)	OWUSU-DARKO ISAAC (PG3943009)	**Signature:……………**	DATE;………..

Certified by:

(Supervisor)	NANA KENA FREMPONG	**Signature:……………**	DATE:………

Certified by:

(Head Of Dept.)	DR. S. K. AMPONSAH	**Signature:……………**	DATE:………

DEDICATION

I dedicate this thesis to my wife Mavis Ansah and two children Elsie Owusu-Agyeiwaa and Sebastian Owusu-Boateng. May God bless them.

ACKNOWLEDGEMENT

First and foremost, a big ‘Thank You’ goes to ‘The King of Kings’ for His divine help, guidance and inspiration. May His name be forever praised.

I am indebted to Nana Kena Frempong, my Supervisor in the KNUST Department of Mathematics. Had it not been his painstaking supervision, this dissertation would have been just a dream. May his time lost be replaced many times over with more blessings.

It is my pleasure to thank my dearly wife Mavis Ansah for her kindly support towards my education in all diverse ways. Her enthusiasms, interest and continual support in this regard would never wane.

I am also grateful to Mr. Takyi Francis for his company, advice and encouragements through our longed trailed suffering studies in both distance and campus life education. May God continue to bless you in life.

I can make along list of the good things people have done for me, but the list could be endless. However, I thank all and sundry for whose support one way or the other helped me proceed in life up to this far. May God bless all of us.

Finally, I cannot hold back my appreciation to my fellow students, friends and love ones, for their concern especially Vincent Tulasi, Duah Dwomour and Frank Bio. I own them a lot of acknowledgements.

ABSRTACT

Education as it's often said is the 'King of all knowledge'. In its yard are high expectations of academic performance which is very dear to the heart of all and sundry. Hence, any variable that triggers the academic performance of students evoke the awareness of all.

The aim of this thesis is to review analyses of "Application of Generalized Estimating Equation (GEE) Models on Academic Performance". There are various statistical and mathematical models employed in the analyses of students' academic performance in different level of schools. In this thesis, we formulate the Generalized Estimating Equation (GEE) model approach to analyse the probable performance of students knowing his 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, Coefficient Estimation of the Study Parameters reveals that, only the geographical location of students is significant and hence affects their academic performance. We recommend that Mathematics Education should be strengthened in the Northern belts regions of Ghana. We also recommend further research to check individual differences existing among the students that may account for differences in Academic performance in Institutions.

TABLE OF CONTENTS

CHAPTER TWO

LITERATURE REVIEW

CHAPTER THREE

METHODOLOGY

CHAPTER FIVE

SUMMARY, CONCLUSSIONS AND RECOMMENDATIONS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABREVIATIONS

AAP	Academic Average Performance
AP	Average Performance
CGPA	Cumulative Grade Point Average
CSPSS	Computerized School Placement And Selection System
CWA	Cumulative Weighted Average
CPM	Covariance Pattern Models
GEE	Generalized Estimating Equation
GES	Ghana Education Service
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Models
IQ	Intelligent Quotient
IRLS	Iterative Regression Least Squares
JHS	Junior High School
KNUST	Kwame Nkrumah University of Science and Technology
KYAE	Kenteyky Adult Education
L/LOC	Location (Geographical Locations)
MAFY	Managed Application Fiscal Year
MRM	Mixed-Effect Regression Model
MLRM	Multiple Linear Regression Model
MML	Marginal Model for Longitudinal (data)
SEB	Socio-Economic Background

SHS	Senior High School
SPSS	Statistical Package for Social Scientists
SSCE	Senior Secondary Certificate Examination
SSS	Senior Secondary School
SWA	Semester Weighted Average
TLM	Teaching And Learning Materials.
UCC	University Of Cape Coast
WAEC	West African Examination Council
WASSCE	West African Senior Secondary School Certificate Examination

CHAPTER ONE

INTRODUCTION

1.0 Overview

This chapter gives a background of the study, statement of the problem, purpose of the study, objectives and hypotheses guiding the study, methodology and significance of the study. In addition to these are limitations as well as organization of the study. The chapter concludes with it summary.

1.1 Background to the Study

Education is a complex process and many factors directly or indirectly affect school outcomes and students academic achievements. As a result, it is difficult to properly define the major factors influencing students' achievement.

Students' academic performance, whether good or bad are accounted for by certain identifiable variables which need to be tested for better confirmation whether truly, this affect students academic performance with respect to their SWA scores. The preliminary investigations of some researchers in their bid to find out whether there exist certain variables affecting academic performance indicate a positive relationship between students' achievement in mathematics and home background variables such as "parents' level of education", mathematics students and their SHS background, gender, fee paying status, loan financial assistance, literacy status of students, choices of students' optional courses, and attitudes of students towards mathematics. Furthermore, the relative effect of the index of socio-demographic factors such as age, gender as well as self-concept affect students' academic performance as well in mathematics, (Papanastasiou, 2002)

In the same way, students “who come from a family in a particular geographical location may perform better and have their SWA highly rated. The mathematics Achievement score of some students with respect to their Semester Weighted Average (SWA) may be attributed to personal attitudes developed towards the study of Mathematics. Moreover, some investigations carried by researchers indicated that students who have positive perceptions or attitudes towards mathematics showed better achievement in both mathematics and science (Kiamanesh, 1997). Whether a student views herself or himself as a strong or weak person in a specific subject may be an important factor in his or her academic achievement. (Stodalsky et al, 1991) mentioned that students develop ideas, feelings and attitudes about school subjects over time and from a variety of sources.

Instruction in school settings provides one important and regularly experienced context in which ideas and perceptions about subject matters as well as other cognitive and affective outcomes can be shaped. The environment that nurtured the child can have influence on the child’s development and willingness to embark on Education. In the cities, the track of educating the child has become highly competitive among neighbours in a particular geographical location. These environmental effects may come from extra-parental influences, such as peer groups and social pressures. When this is realized, it affects children preparedness to meet the competition ahead, and hence affect his or academic performance when they reach the Universities, (Papanastasiou, 2002).

Researchers (Frize et al., 1983; Weiner, 1985) showed that attributions influence students’ achievement. Students often attribute their outcomes to variables like hard work, good luck and natural talent. Even though students may attribute their failure or success to the afore-mentioned variables, the efforts that they make in order to learn mathematics at school or do assignments probably have an effect on their academic performance achievement.

There is therefore the need to investigate into an analyses of students performance in Kwame Nkrumah University of Science and Technology (KNUST) similar to abundant research carried out in this field to find out whether or not, a significant effect of certain variables such as age, gender, geographical location of students as well as Ghana Education Service (GES) graded level of Senior High School (SHS) students attended have adverse effect on students academic performance in mathematics. These concepts in various multiple variables are related to KNUST Mathematics students' Semester Weighted Average (SWA) score which the study seeks to investigate. Some of these multiple variables may be student's gender, age, and geographical location of students as well as graded level of formal school attended and perception of students towards mathematics studies.

The use of Semester Weighted Average (SWA) is however, 'an academic assessment tool used for striking the academic average performance of students achieved in his or her placed institution'.

1.2 Statement of the Problem

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.

1.3 Objectives

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 achieved in the University.

1.4 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.

1.5 Justification.

The study would afford students the opportunity to be aware of certain existing variables that may have an impact on their academic performance with respect SWAs. Information gathered from results would inform students about their academic performance based on certain factors. This would help fill the existing gap in the research carried out in Ghanaian Universities in this area. In addition, it could pave the way for more comprehensive research on the comparison of national and international research findings on factors affecting students' academic performance.

The study would how ever be useful to authorities in the university. Authorities would be alerted of these significant variables that plays restrictions on students academic.

The study would equally be helpful to parents to understand the underlining cause and effects of their children's inability to meet learnable skills to successful semester grades.

The University as a whole would find the study relevant in keeping tracks of students record in successful and failure records of grades (SWA), and embark on further research on this in order to find a plausible solution to the impending problem.

1.6 Limitations

Aside indisputable constraints such as financial and time, other uncontrollable constriction beyond the control of the researcher may impinge restrictions on the conclusion of the study. Some of these factors proposed to conflict the study may be data-gathering instruments. In as much as the researcher would minimized occurrence of biasedness, unwillingness of respondents to disclose a validated response to the given questionnaires could falter the true results of the analyses.

1.7 Organization of the study

Chapter one is made up of introduction, which comprises the background of the study, purpose of the study, statement of the problem, research question and hypotheses, significance of the study, and limitations. Chapter two highlights on review of literature of ideas of different authors whose findings have been defined in relation to the topic under study. Chapter three focuses on methodological review in the light of Mathematical Statistics tools that are relevant to the analyses of the various data gathered. Basically, the study seeks to use GEE model for the analyses. Chapter four deals with data analyses. In the same way, chapter five consists of summary, conclusion and recommendations.

The project report however ends with references and appendices in supportive to the researcher's investigation.

1.8 Chapter Summary

The chapter gave an introduction to the thesis report highlighting on issues relating to background of the study, statement of the problem, purpose of the study, objectives and hypotheses guiding the study, methodology and significance of the study. In addition to these are limitations as well as organization of the study. The chapter concludes with this summary

CHAPTER TWO

LITERATURE REVIEW

2.0 Overview:

In this section, there is a review of the work of several authors concerning concept definitions and various researches done to uncover the academic achievement of students in the tertiary institutions. Researches, empirical work and authors' opinion are looked at. Below are the focuses of the review.

- The Concept Definition of academic performance, SWA:
- Study variables affecting students SWA
- Academic Performance In Universities
- The Grade Level and Background Of Students' Secondary School Education
- Gender issues versus Mathematics performance.

2.1 The Concept Definition of academic performance, SWA:

Kenteyky Adult Education [KYAE] report on Managed Application Fiscal Year (MAFY, 2009-10) defines academic performance and retention as "a process where a student's success in school is measured to determine how they stand up to others in the same areas. Academic performance refers to how students deal with their studies and how they cope with or accomplish different tasks given to them by their teachers. Academic performance is the ability to study and remember facts and being able to communicate what is learnt in successful manner. Most people know that academic performance generally refers to how well a student is accomplishing his or her tasks and studies, but there are quite a number of factors that determine the level and quality of students' academic performance”. Pg 15 To them, Academic Performance is measured in terms of “the percent of enrolled students completing educational levels”. Program’s Academic Performance is the ratio of total number of students completing educational levels to the total number of students enrolled in educational programme.

Thus, the Academic Performance Index (API), is a process used in California to determine the performance and growth of students in the world of academics to learn educational facts.

According to Bell (2010) contribution in defining the meaning of academic performance in *eHow* (A *eHow* Contributor), he commented that, *in educational* institutions, success is measured by academic performance, or how well a student meets standards set out by local government and the institution itself. As career competition grows ever fiercer in the working world, the importance of students doing well in school has caught the attention of parents, legislators and government, as well as education departments alike. Although education is not the only road to success in the working world, much effort is made to identify, evaluate, track and encourage the progress of students in schools. Parents care about their child's academic performance because they believe good academic results will provide more career choices and job security. Schools, though invested in fostering good academic habits for the same reason, are also often influenced by concerns about the school's reputation and the possibility of monetary aid from government institutions, which can hinge on the overall academic performance of the school. State and federal departments of education are charged with improving schools, and so devise methods of measuring success in order to create plans for improvement.

In the past (from history), academic performance was often measured more by ear than today. Teachers' observations made up the bulk of the assessment, and today's summation or numerical method of determining how well a student is performing is a fairly recent invention. Grading systems came into existence in America in the late Victorian period, and were initially criticized due to high subjectivity. Different teachers valued different aspects of learning more highly than others, and although some standardization was attempted in order to make the system fairer, the problem continued. Today, changes have been made to incorporate differentiation for individual students' abilities, and exploration of alternate methods of measuring performance is ongoing. Some Universities use the Semester Weighted Average (SWA) system whilst others employ the Cumulative Grade Point Average (CGPA). The use of SWA or CGPA is an academic assessment too used for striking the academic average performance of students achieved in his or her placed institution (Bell, 2010).

Functionally, the tracking of academic performance fulfills a number of purposes. Areas of achievement and failure in a student's academic career need to be evaluated in order to foster improvement and make full use of the learning process. Results provide a framework for talking about how students fare in school, and a constant standard to which all students are held. Performance results also allow students to be ranked and sorted on a scale that is numerically obvious, minimizing complaints by holding teachers and schools accountable for the components of each and every grade.

Considering academic Features, Performance in school and Universities is evaluated in a number of ways. For regular grading, students demonstrate their knowledge by taking written and oral tests, performing presentations, turning in homework and participating in class activities and discussions. Teachers evaluate in the form of letter or number grades and side notes, to describe how well a student has done. At the state level, students are evaluated by their performance on standardized tests geared toward specific ages and based on a set of achievements students in each age group are expected to meet.

In considerations to the said topic under discussion, the subjectivity of academic performance evaluation has lessened in recent years, but it has not been totally eliminated. It may not be possible to fully remove subjectivity from the current evaluation methods, since most are biased toward students that respond best to traditional teaching methods. Standardized testing is best responded to by students that excel in reading, mathematics and test-taking, a skill that is not in itself indicative of academic worth. The tests reward visual learners, and give no chance for kinesthetic or auditory learners to show their abilities. The standardized test fails to recognize students with learning and physical disabilities that do not allow them to complete the test in the same manner or amount of time as other students. Evaluations from classroom teachers, though they give the most detailed information, may still retain bias if individual differentiation and learning styles have not been taken into account.

2.2 A REVIEW OF THE STUDY VARIABLES

There are various rich literature on the factors that influence a students' performance in their first year of study. This study is however delimited to the following variables:

2.2.1 Gender issues versus Mathematics performance

According to (Evans, 1999), the gender of the student is also important. “Overall, females generally perform better than males, but there are exceptions in some disciplines”

All of the research reviews support the hypothesis that student performance depends on different factor including gender (sex of the person). The findings of research studies focused that student performance is affected by different factors such as learning abilities because new paradigm about learning assumes that all students can and should learn at higher levels but it should not be considered as constraint because there are other factors like gender, sex that can affect student’s performance. (Hansen & Joe, 2000).

The performance of students on the module is not affected by such factors as age, sex and place of residence but is associated with qualification in quantitative subjects as evidenced in (Soyibo et al., 1998) report findings.

(Winston et al., 2002) focused on student’s impatience (his time-discount behavior) that influences his own academic performance. Between male and females are existence of patience perseverance and tolerance in meeting learning standards especially on Mathematics. (Goethe, 2001) found out that weak students do better when grouped with other weak students and are mixing male and female students together in learning. As implied by (Zajonca, 1976) analysis of older siblings, it shows that students’ performance improves if they are with the students of their own kind irrespective of gender and age. There are often different results by gender, as in

(Hoxby, 2000) result. (Sacerdote, 2001) found out that grades are higher when students have unusually academically strong room mates.

The results of (Zimmerman, 2001) were somewhat contradictory to (Goethe, 2001) results but again it proved that students performance depends on number of different factors, it says that weak peers might reduce the grades of middling with strong students. (Gur et al., 1977) explained that some of the practices adopted by college administration in higher education like residential colleges or organized study groups also help to increases performance.

2.2.2 Age determinant variables

The age of a student has also been found to be important. Younger students are more likely to complete qualifications than older students in actual terms, but older students generally outperform their younger counterparts when controlling for full-time and part-time study status (Scott & Smart, 2005). Age, though, is correlated with taking a gap-year. That is, a student age, whether young or old (perhaps was admitted on mature students entrance exams consideration) can affect students academic performance There are two other factors that have a large impact on tertiary success; the level of study (certificate, diploma, bachelors, post-graduate) and study load that is coupled with students age limit.

Students who take some time off before starting tertiary study will generally be older in their first year of study than a student who progresses directly to tertiary study after leaving school. Students who take a gap-year out-perform students who progress directly (Birch & Miller, 2007). The age of a student is also correlated with maturity and motivation, which has been shown to be a good predictor of academic performance (Evans, 1999).

2.2.3 Geographical Location and background of students:

(Birch &Miller, 2007) an Australian study found that students from middle-level socioeconomic communities performed better than lower socio-economic community students of the same ability level, who in turn performed slightly better than higher socio-economic community students of the same ability level. They suggested this was because higher socio-economic families disproportionately send their children to non-government schools. Studies have shown that nongovernment school students in Australia do not perform as well at university as government school students when school achievement is controlled for. (Birch & Miller, 2007).

Ethnic group is also sometimes associated with academic achievement (Evans, 1999). The few Australian studies that have examined the impact of ethnic background on grades have indicated only a small impact on academic performance (Birch & Miller, 2004). Australian students from non-English speaking backgrounds have been found to have slightly higher grades than students from English speaking backgrounds. This was attributed to there being a greater motivation to study at university due to cultural factors that place a premium on education (Birch & Miller, 2004). (Scott & Smart, 2005) showed that in New Zealand, Maori and Pasifika had the lowest degree-completion rates, even after adjusting for various demographic, geographical location and study related factors. Some studies have found that gaining entry to tertiary studies and students' persistence and success in tertiary study, are related to socio-economic and demographic status (Evans, 1999).

Some studies suggest that there are factors that affect academic performance that are specific to minority ethnic groups and locational background of students, and relate to the interaction between the student and the institution. Among those factors are isolation, alienation and lack of support (Allen, 1992), and perceptions of prejudice and discrimination (Nora & Cabrera, 1996). Although these studies refer in the main to American students, institutional factors may also be playing a part in Ghanaian context.

2.2.4 Grade Level of SHS Attended:

The factor that is most correlated with first year tertiary students' performance is their previous academic performance in school. Students, who perform well in secondary school, or even primary school, do well at university (Birch & Miller, 2004).

In Ghana, GES has categorized SHS according to availability of facilities, geographical location, subjects offered and vacancies available to allow candidates spread their choices so as to increase their chances of being placed through the Computerized School Selection and Placement System (CSSPS). The intention is to enhance students' enrollment in schools within their own vicinity or geographical location. Schools had been put into six categories of Public Senior High Schools in 'A,' to 'D', Public Technical/Vocational, Institutions under category 'T' and Private Senior High Schools and Technical Vocational Institutes in category 'P'. A Candidate is expected to choose a total of six schools, stressing that candidates would be allowed to choose only one school from category 'A' and a maximum of two schools from category 'B'.

Irrespective of students' choices of school preference, the issue of entering the school with certain italic characteristics which can correlate, whether or not with their academic performance in their Senior High School placed are however not considered. (GES, 2009), www.modernghana.com/.../computer-selection-and-placement-system-*reviewed.html*

Ghanaian schools apart from partitioning them into stages and levels can also be put into day and boarding schedules based on facilities available or the huge number intake. Schools in the revolutionary days used to run day-school except higher levels in the child education. Nowadays, boarding is realized even from the crèche, nursery and primary, Junior High School (JHS), Senior High School (SHS) and other tertiary institutions. Some students begin their lifetime schooling through the forefront of boarding and hardly take in when placed in the day school. Such entering characteristic, when realized can adversely affect students focus in learning relative to academic performance especially when they reach tertiary level.

The qualification a student uses to enroll in University also plays vital role in the determination of academic performance. Other studies have shown links between students' domain-specific knowledge acquired and students' previous school and their motivation to study a subject, which also relates to their course preference and academic preparedness (Evans, 1999). The link is however strongest for science disciplines of which in a sense is an entry requirement into KNUST

2.3 Academic Performance In Universities

As already mentioned in the afore discussion, success in every school or institution is measured by academic performance, or how well a student meets standards set out by local government and the institution itself especially in educational institutions such as the university. Certain factors, perhaps beyond control, can impinge restriction of students' academic performance in getting a high SWA. For example, Studies exploring the relationship between socio-demographic factor and academic performance have produced mixed results. A study by (Schutte et al., 1998) found that scores on a self-report measure of socio-demographic factor completed at the beginning of the academic year significantly predicted grade point average at the end of the year. In a study by (Pettijohn & Parker, 2002), there was a small, but significant relationship between academic success, as measured by grade point average, and three out of the five factors within the utilized emotional intelligence scale utilizing the (Goleman, 1998) scale.

Understanding the causes and effects of various ways in which performance is altered is an important element of Intelligence Quotient (IQ). (Rode et al., 2007) continued by including the research of (Mayer & Salovey, 1997): individuals with a high level of intelligence are able to direct positive ambitions to uphold the energy needed for high performance over long periods of time in University learning environment. Thus, (Rode et al., 2007) reasoned that individuals with high emotional intelligence would perform better academically. Despite their prediction, emotional intelligence was not significantly associated with grade point average of what others term Semester Weighted Average (SWA). However, they did find an interaction of intelligence with conscientiousness explained unique variance in academic performance (cumulative GPA), as well as public speaking and group behavior effectiveness.

2.4 LONGITUDINAL DATA AND MARGINAL MODELLING

In the 1980s, alongside development of MRMs and CPMs for incomplete longitudinal data, generalized estimating equations (GEE) models were developed (Liang and Zeger, 1986). Essentially, GEE models extend generalized linear models (GLMs) for the situation of correlated data. Thus, this class of models has become very popular especially for analysis of categorical and count outcomes, though they can be used for continuous outcomes as well. One difference between GEE models and MRMs and CPMs is that it uses quasi-likelihood estimation, and so the full likelihood of the data is not specified. GEE models are termed marginal models, and they model the regression of y on X and the within-subject dependence (***i.e.,*** the association parameters) separately.

As noted in (Fitzmaurice et al., 2004), "the term marginal in this context indicates that the model for the mean response depends only on the covariates of interest, and not on any random effects or previous responses." In statistical terms, these two parameter vectors are assumed to be orthogonal to each other; GEEl is the class of models that is most commonly found in statistical software implementations, some of which are reviewed in (Horton & Lipsitz, 1999).

Subsequent to the development of the GEEl class of models, GEE2 models were developed that do not make this separation of the regression and association parameters. In other words, GEE2 does not assume orthogonality of these parameter vectors. GEE models will refer only to the GEEl class of models. GEE models have important differences from MRMs, and these are well-described by several authors (Burton et al., 1998; Diggle et al.; Zeger et al., 1986).

2.5 GENERALIZED LINEAR MODELS (GLMS)

Before describing GEE models, it is useful to review generalized linear models (GLMs), since GEE models can be viewed as an extension of GLMs to the case of correlated data. GLMs represent a class of models that are used to fit fixed effects regression models to normal and non-normal data. (McCullagh & Nelder, 1989) describe this class of models in great detail and point out that the term “generalized linear model” is due to (Nelder & Wedderbum, 1972), who indicated how linearity could be exploited to unify several diverse statistical techniques. The essential idea is to treat many types of regression models, which differ primarily in terms of the type of dependent variable they model, as special cases of a single family of models. The dependent variable is assumed to come from the class of distributions known as the exponential family, and common GLM family members include linear regression for normally distributed dependent variables, logistic regression for dichotomous dependent variables, and Poisson regression for counts.

Under the identity link, the expected value of the dependent variable is simply a linear function of the explanatory variables multiplied by their regression coefficients. For dichotomous outcomes, logistic regression is applied (Hosmer & Lemeshaw, 2000)

GLMs are fixed effects models which assume that all observations are independent of each other. Thus, they are not generally appropriate for analysis of longitudinal data. However, they can be extended to account for the correlation inherent in longitudinal data, and this is what Liang and Zeger did in developing GEE models, (Liang & Zegger, 1986). It should be noted that other models may be described on mixed-effects logistic and Poisson regression. These, also represent generalizations of GLMs by including random effects, and thus represent generalized linear mixed models (GLMMs). GEE is a different kind of generalization of GLM than that provided by GLMMs.

2.6 Chapter summary:

This present study reviews other related literature on the topic with respect to the variables under consideration and looks at a number of the factors explored in the literature researched as summarized above. Academic Performance (AP) achievement at school could be related to age, gender, GES grade level of students SHS background geographical location of the student (including within ethnic group distinctions). The socio-economic rating of the last secondary school attended is included – this is a proxy for the socio-economic standing of the student's Academic performance of first-year bachelors students at university. The timing of the progression to tertiary study — whether the student went directly after leaving school, or took a year off — is also included. Gender, whether the student studied intra- or extramurally, and whether the student studied full-time or part-time, are considered, and controlled for. Age is not considered separately because this study is restricted to a narrow age range, and within that range, age is correlated with the timing of progression.

CHAPTER THREE
METHODOLOGY.

3.0 Research Design:

The research study is descriptive study and the main design used was descriptive sample survey, which is mainly concern with the description of some existing phenomenon about the academic performance characteristics of students in Universities relative to their academic average score in SWA. The researcher chose this for the study because he considered it to be the most appropriate one for the investigation. This is to help make a generalization based on the fraction of the population sampled in the case study.

However the mathematical Methodology that was employed for the study was Generalized Estimating Equation (GEE) model e.i. a Marginal Model Longitudinal (MML) data approach since the variables under study are Multivariate with two or more dependents and independents variables.

3.1 Population:

The target population involved a census population of all Level 400 students in the KNUST Mathematics department in the Ashanti region of Ghana. The total number of Students' population mount to approximately 139. Out of these, the completed students SWA track records used for the study were 126 with 97: 28 male and female ratio respectively. The researcher deems it appropriate to target population sample to only KNUST Mathematics students as a case study. Among the Level 400 students, 105 are males whilst the rest of 34 are female offering Mathematics.

3.2 Sample and Sampling Procedure:

The entire population of MATH IV (2010/2011 academic year was obtained from the exams office. 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.

3.3 Research Instrument:

(Amadehe, 2002) and (Kerlinger, 1974), concluded that questionnaire is one of the best instruments procedure in descriptive design in research. “Questionnaire, when widely used for collecting data in educational research and if developed to answer questions is very effective in securing factual information about practice and conditions at which the respondents are presumed to have knowledge and for inquiring into the opinions and attitudes of the subject under study”. (Kerlinger, 1974).

Thus, questionnaire was used in addition to the existing data of students SWA record to inquire opinions from students on their perceptive view about factors (variables) they think affect their performance in Mathematics. There were 15 question items consisting of close and open items, structured and unstructured ones. The preamble of the questionnaire dealt with instructional alertness to students. The questionnaire were put into four (4) sections from A to D with each section having a treatment of inquiring about at least a study variable of students in correspondence to the research question/hypotheses.

Questionnaires were organized to solicit information from students on the extent to which the researcher variables of interest affect their Academic Performance (A. P.). Respondents were instructed to tick [√] the options from the appropriate boxes the one that deemed supportive to each statement as seen from the appendix one (A) of the project report.

3.4 Administration of Instrument and Data Collection Procedure:

The questionnaire was the main instrument administered. Before the administration of the instrument, it was given a face validation by the supervisor of this study. It was then piloted in a different Department in the same University. This was to find out if it could demand the needed information in correspondence to factors affecting their performance.

All questionnaires were successfully collected constituting 100% of the total samples. The questions were however answered well and contributed immensely towards the statistical analyses of the project report.

3.5 Data Analyses:

The data collected were edited for consistency. The research results were presented in the form of frequency distribution tables which was later translated into matrices computational form to calculate determinants, cofactors, inverse and further solutions. Display of charts and graphical displays showing the frequencies of responses to the research questions and SWA were shown. This played a contributive factor in description of the observed results using simple percentages (%) and analyses of GEE family of models as discussed in chapter four (4) hereafter.

Since the study sought to find validate significant difference between students academic performance and existing variables (age, gender, geographical location and graded level of former SHS status student attended), Generalized Estimating Equation family models (comprising GEE-Independent, Exchangeable, Autocorrelation and Unstructured models) were supportive in the analyses. Microsoft Excel version 2007, SPSS version 16, Minitab version 14 and SAS version 9.1 statistical software were used to analyze the data and further statistical analyses and conclusion made leading to the outline and conclusion of the project report.

3.6 MARGINAL MODELS FOR LONGITUDINAL DATA

Generalized estimating equations (GEE) models were developed by Liang and Zeger in the 1980s, alongside development of MRMs and CPMs for incomplete longitudinal data, (Zeger and Liang, 1986). Essentially, GEE models extend generalized linear models (GLMs) for the situation of correlated data. Thus, this class of models has become very popular especially for analysis of categorical and count outcomes, though they can be used for continuous outcomes as well. "One difference between GEE models and MRMs and CPMs is that it uses quasi-likelihood estimation. GEE models are termed marginal models, and they model the regression of y on X and the within-subject dependence (***i.e.,*** the association parameters) separately. As noted in (Fitzmaurice et al., 2004), "the term marginal in this context indicates that the model for the mean response depends only on the covariates of interest, and not on any random effects or previous responses." (Hedeker & Roben, 2006)

(Diggle et al., 2002) introduce the topic of longitudinal data. Traditional data analysis depends on the assumption of independence. Data are collected and analyzed, and inferences are made. Longitudinal data must address issues of temporal correlation, along with small numbers of independent (usually) subjects. Methods developed for dealing with these types of data are borrowed from traditional data analysis of independent data, and time series methods.

Correlated data can arise from situations such as

1. longitudinal studies, in which multiple measurements are taken on the same subject at different points in time (e.g. students SWA scores across seven semesters in their four year academic track records)
2. clustering, where measurements are taken on subjects that share a common category or characteristic that leads to correlation.

The following sections describe the models developed to analyze longitudinal data such as students AP achievements from seven (7) semesters in relation to their boidata response (such as age and gender), geographical location and graded level of former school.

3.7.1 GENERALIZED LINEAR MODELS (GLMS)

GLMs were developed as an extension to linear models, to allow for more complex relationships between the response and the explanatory variables, e.g. binary or count data. Generalized Linear Models have three main components: a family or distribution (the exponential family, including all the standard distributions used in GLMs), a linear predictor and a link function. Instead of having

$E(Y_{it}) = \mu_i = x_i'\beta$ we now have

$$E(Y_{it}) = \mu_i \text{ and } g(\mu_i) = \eta_i = x_i'\beta \tag{3.1}$$

where *g(.)* is a monotone link function. The main assumptions involved with GLMs are as follows (Hardin & Hilbe, 2001):

- that the *Yit*'s are independent (i.e., uncorrelated),
- that the variance function $V(\mu_i)$ is correctly specified,

- that the dispersion parameter $\hat{\varphi}$ is correctly specified (i.e., is equal to one for Binomial and Poisson data), and,
- that the link functions is correctly specified.

Linear models are a special case, when the link function is the identity link and the distribution is normal. For the Poisson distribution, the natural or canonical link function is the log link. GLMs do not assume constant variance, but assume that there is a known relationship between the mean and variance. They also assume linearity on the scale of the link function. GLMs solve the problem of non-normality and non-constant variance. Before describing GEE models, it is useful to review generalized linear models (GLMs), since GEE models can be viewed as an extension of GLMs to the case of correlated data. GLMs represent a class of models that are used to fit fixed effects regression models to normal and non-normal data. (McCullagh & Nelder, 1989) describe this class of models in great detail and point out that the term "generalized linear model" is due to (Nelder & Wedderbum, 1972), who indicated how linearity could be exploited to unify several diverse statistical techniques. The essential idea is to treat many types of regression models, which differ primarily in terms of the type of dependent variable they model, as special cases of a single family of models. The dependent variable is assumed to come from the class of distributions known as the exponential family, and common GLM family members include linear regression for normally distributed dependent variables, logistic regression for dichotomous dependent variables, and Poisson regression for counts. There are three specifications in a GLM. First, the linear predictor, denoted as η, of a GLM is of the form

$$\eta_i = x_i' \beta \qquad (3.2)$$

where x_i is the vector of explanatory variables, or covariates, for subject *i* with fixed effects β. This first step indicates a linear predictor η_i which is based on covariates x_i and regression coefficients β. The covariates in x_i can include continuous repressors', dummy variables, interactions, polynomials, etc. Then, a link function $g(.)$ is specified which converts the expected value μ of the outcome variable y (i.e., $\mu_i = E[y_i]$) to the linear predictor η.

$$g(\mu_i) = x_i'\beta = \eta_i \qquad (3.3)$$

For example, in ordinary multiple regressions, the link function is called the identity link since

$$g(\mu_i) = \mu_i \quad \text{and so} \quad \mu_i = \eta_i, \text{ or}$$

$$E(y_i) = x_i'\beta = \mu_i = \eta_i \qquad (3.4)$$

Under the identity link, the expected value of the dependent variable is simply a linear function of the explanatory variables multiplied by their regression coefficients. For dichotomous outcomes, logistic regression (Hosmer & Lemeshaw, 2001) is a popular choice for analysis. This model is written as

$$log\left[\frac{P(y_i=1)}{1-P(y_i=1)}\right] = x_i'\beta \qquad (3.5)$$

where y takes on values of 0 or 1. Since $P(y_i = 1) = E(y_i) = \mu_i$ in this case, we see that it is the logit link $g(\mu_i) = \log[\frac{\mu_i}{(1-\mu_i)}]$ which relates the expected value of the outcome variable to the linear predictor.

Similarly, the Poisson regression model (Cameron & Trivedi, 1998), which is used to model count data, is written as the probability distribution given as $f(y;u) = \frac{e^{-\mu}\mu^{y}}{y!}$ for $y \geq 0$ or written in a form to make it comparable with Equation (3.5), the log-likelihood becomes

$$\log(l) = \frac{y\log(\mu)}{1} - \log(y!)$$

The denominator of 1 refers to the dispersion parameter φ of Equation 3.7

$$\mu_i = Exp(x_i'\beta) \qquad \text{Or}$$

$$log(\mu_i) = x_i'\beta, \tag{3.6}$$

which shows that it is the log link $g(\mu_i)= \log\mu_i$ that is used for Poisson regression. So far, we've specified what the covariates are and how they relate to the expectation of the dependent variable. In a GLM, we additionally need to specify the form of the conditional variance of y, given the covariates.

This is done as

$$V(y_i)= \emptyset v(\mu_i) \quad = \varphi v(\mu_i) \tag{3.7}$$

Where $v(\mu_i)$ is a known variance function and φ is a scale parameter that may be known or estimated. For example, for ordinary multiple regression $v(\mu_i) = 1$ and φ would represent the error variance (*i.e.,* φ represents the variance of the conditional normal distribution of y given *x*) which is estimated. For a dichotomous outcome, the Bernoulli distribution specifies

$$v(\mu_i) = \mu_i(1 - \mu_i) \tag{3.8}$$

and φ is typically not estimated but set to 1 in the ordinary GLM. An exception is for models that allow over- or under-dispersion, in which case φ is estimated. For a count outcome, the Poisson distribution specifies that the mean equals the variance, and so $v(\mu_i) = \mu_i$, where again φ is set to 1 (*i.* e., it is not estimated) in the usual GLM. We conclude that the link function and variance specification usually depend on the distribution of the outcome variable y. With these GLM specifications, one can estimate the regression coefficients β by solving the estimating equation

$$U(\beta) = \sum_{i=1}^{N}\left(\frac{\partial\mu_i}{\partial\beta}\right)'(V(y_i))^{-1}[y_i - \mu_i] = 0 \tag{3.9}$$

For example, in the ordinary multiple regression models, we get the usual

$$U(\beta) = \sum_{i=1}^{N} \boldsymbol{x_i}\,(y_i - x_i'\,\beta,)$$

(3.10)

for solution of the regression coefficients β.

As noted by (Wedderturn, 1974), the above estimating equation (3.8) depends only on the mean and variance of y, and therefore the precise distributional form for y is not necessary for estimation of the regression coefficients β. In this case, solution of this estimating equation provides what are called “quasi-likelihood” estimates.

3.7.2 Quasi-Likelihood

All of the above GLM theory depends on choosing a distributional form for the data (e.g. Binomial, Gaussian or Poisson) and deriving a likelihood function with its resulting theoretical properties. Often, though, the observed data do not correspond to any distribution exactly, and so we cannot rely on the maximum- likelihood function for estimation. For this reason, an extension was developed - the quasi-likelihood function, where only the relationship between the mean and the variance of the observations needs to be specified. The quasi-likelihood function $Q\ (yi;\ \mu_i)$ is defined as:

$$\frac{\partial Q(y_i, u)}{\partial ui} = \frac{y_i - u_i}{V(u_i)}$$

(3.11)

or equivalently

$$\int^{u_i} \frac{y_i - u_i}{V(u_i)}\, d_i + f(y_i).$$

(Wedderburn, 1974) describes how estimation using maximum quasi-likelihood is directly equivalent to estimating using maximum likelihood, without having to rely on choosing the correct distribution for the observed data. (Nelder, 2000) acknowledges that one of the most important quasi-likelihoods is that for the over dispersed Poisson distribution. If the link and variance correspond to a particular member of the exponential family, then the quasi-likelihood is equal to the likelihood proper.

3.8. 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

$$\eta_{ij} = x'_{ij}\beta \tag{3.12}$$

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

$$g(\mu_{ij}) = \eta_{ij} \tag{3.13}$$

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,

$$V(\mu_{ij}) = \varphi v(\mu_{ij}) \tag{3.14}$$

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

3.8.1 Specifying the Correlation Matrix

An important aspect of the GEE is specifying the form of the correlation matrix, $R_i(a) = I$, a *n* x *n* identity matrix. According to (Liang & Zenger, 1986), the GEE approach yields a consistent estimator of β's even when R_i is misspecified. For this reason, an independence model is often used when the choice of R_i is not obvious. The most commonly used working correlation are:

GEE Independence model: Independent working correlation assumes that there is no correlation within the clusters of students' SWA scores and the model becomes equivalent to standard normal regression. Independence assumes that there is no correlation within the clusters and the model becomes equivalent to standard normal regression with the identity matrix I

$$Corr\ (y_{ij}, y_{i,j+k}) = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} = I$$

GEE Exchangeable (Compound symmetry): 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)

$$Corr\ (y_{ij}, y_{ii}) = \begin{cases} 1, & j = k \\ \rho, & j \neq k \end{cases}, \begin{pmatrix} 1 & \rho & \rho & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & \rho & \rho & 1 \end{pmatrix}$$

GEE First order Auto-regressive (AR-1): Autoregressive weights the correlation within clusters by their separated time and hence correlation coefficients diminish for further distances. Similar to exchangeable model, it requires only one estimated parameter. For application of GEE models, one assumes that there are a fixed number of time-points *n* that subjects are measured at.

Where

$$Corr\,(y_{ij}, y_{i+k}) = \rho^{k}, \begin{pmatrix} 1 & \rho & \rho^{2} & \rho^{3} & \rho^{4} & \rho^{5} & \rho^{6} \\ \rho & 1 & \rho & \rho^{2} & \rho^{3} & \rho^{4} & \rho^{5} \\ \rho^{2} & \rho & 1 & \rho & \rho^{2} & \rho^{3} & \rho^{4} \\ \rho^{3} & \rho^{2} & \rho & 1 & \rho & \rho^{2} & \rho^{3} \\ \rho^{4} & \rho^{3} & \rho^{2} & \rho & 1 & \rho & \rho^{2} \\ \rho^{5} & \rho^{4} & \rho^{3} & \rho^{2} & \rho & 1 & \rho \\ \rho^{6} & \rho^{5} & \rho^{4} & \rho^{3} & \rho^{2} & \rho & 1 \end{pmatrix}$$

Unstructured/Unspecified GEE model: In unstructured working correlation structure specification in GEE modeling, we assume different correlations between any two measurements on SWAs for every students. No constraints are placed on the correlations. Every element of the correlation matrix is estimated separately.

$$Corr(y_{ij}, y_{i,k}) = \begin{cases} 1, & j = k \\ \rho_{jk}, & j \neq k \end{cases}, \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} & \rho_{15} & \rho_{16} & \rho_{17} \\ \rho_{21} & 1 & \rho_{23} & \rho_{24} & \rho_{25} & \rho_{26} & \rho_{27} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} & \rho_{35} & \rho_{36} & \rho_{37} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 & \rho_{45} & \rho_{46} & \rho_{47} \\ \rho_{51} & \rho_{52} & \rho_{53} & \rho_{54} & 1 & \rho_{56} & \rho_{57} \\ \rho_{61} & \rho_{62} & \rho_{63} & \rho_{64} & \rho_{65} & 1 & \rho_{67} \\ \rho_{71} & \rho_{72} & \rho_{73} & \rho_{74} & \rho_{75} & \rho_{76} & 1 \end{pmatrix} \quad (3.16)$$

(Zeger et al., 1988) also demonstrated that β obtained under the independence model is relatively efficient. The additional specification in a GEE model is for the “working” correlation structure of the repeated measures. This working correlation matrix is of size $n_i \times n_i$ because one assumes that there are a fixed number of time points n_i that subjects are measured at.

A given subject does not have to be measured at all n time points; each individual's correlation matrix R_i is of size $n_i \times n_i$ with the appropriate rows and columns removed if $n_i < n$. It is assumed that the correlation matrix R_i depends on a vector of association parameters denoted α. Examples of the various working correlation structures below will make this notion more concrete. These parameters α are assumed to be the same for all subjects. They represent the average dependence among the repeated observations across subjects. The simplest form is that of independence, namely $R_i(\alpha) = I$, which is '$n_i \times n_i$' identity matrix. This form is equivalent to assuming that the longitudinal data are not correlated.

The next simplest structure is to assume that all of the correlations in R_i are the same, or "exchangeable." This exchangeable structure specifies that $R_i(a)=\rho$, namely that all of the correlations are equal.

Another useful one parameter model for longitudinal data is the AR (1) structure, namely, $R_i(a)=\rho^{/j-j/}$. Here, the within-subject correlation over time is an exponential function of the lag. For Toeplitz structure $R_i(a)= \rho_{/j-j/}$, if $j - j^1 \leq m$. and $R_i(a) = 0$ if $j - j > m$. Where the fullest structure is $m = n-1$.

3.8.2 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

$$V(a) = \varphi A_i^{\frac{1}{2}} R_i(a) A_i^{\frac{1}{2}} \tag{3.17}$$

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

$$V(a) = \varphi R_i(a) \tag{3.18}$$

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

$$\sum_{i=1}^{N} D_i'[V(\hat{a})]^{-1}(y_i - \mu_i) = 0 \tag{3.19}$$

Where $\hat{a}$ is a consistent estimate of a and $D_i = \left(\frac{\partial \mu_i}{\partial \beta}\right)$ and hence, equation (3.18) becomes

$$\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right) (V(\hat{a}))^{-1}[y_i - \mu_i] = 0 \tag{3.20}$$

This is an extension of the estimating equation for β in any GLM, which is given in (3.19). Thus, the GEE solution can be seen as a natural generalization of the GLM solution for correlated data.

As an example, in the normal case, for equation (3.19), i.e

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

$$\mu_i = X_i \beta \tag{3.21}$$

$$D_i = X_i$$

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

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

$$\beta = \left[\sum_{i=1}^{N} X_i'[R_i(\hat{a})]^{-1}X_i\right]^{-1}\left[\sum_{i=1}^{N} X_i'[R_i(\hat{a})]^{-1}y_i\right] \qquad (3.22)$$

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 $R_i(a)$ 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

$$r_{ij} = \frac{(y_{ij}-\hat{\mu}_{ij})}{\sqrt{[V(\hat{a})_{jj})}} \qquad (3.23)$$

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 $\hat{\beta}$.

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

$$V(\hat{\beta}) = \left[\sum_{i=1}^{N} X_i' \hat{V}_i^{-1^{-1}} X_i\right]^{-1} \quad (3.24)$$

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

$$V(\hat{\beta}) = \sum_{i}^{N} M_0^{-1} M_1 M_0^{-1} \quad (3.25)$$

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

$$M_0 = \left[\sum_{i=1}^{N} D_i' (\hat{V}_i^{-1}) D_i\right]^{-1}$$

$$M_1 = \sum_{i=1}^{N} D_i' \hat{V}_i^{-1} (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)' \hat{V}_i^{-1} D_i$$

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

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

(3.26)

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

3.8.3 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

$$x^2 = \hat{\beta}'C'(CV(\hat{\beta})C')'C\hat{\beta} \tag{3.27}$$

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: $\beta_q \neq \beta_p$

8.9 SUMMARY

This chapter has described the methodological approach to the Thesis report. Emphasis was laid on the research design, population and sampling techniques as well as the GEE approach for longitudinal data analysis. This approach has several features which makes it particularly useful and popular. Because it is a generalization of GLM, many types of dependent variables can be accommodated within the GEE family of models. As noted, GEE can also be applied to other types of outcomes such as continuous, counts, ordinal, or nominal dependent variables. For all of these, GEE provides regression estimates that are "population-averaged" rather than the "subject specific" estimates of the mixed-effects regression models. These two types of estimates agree in terms of scale only for continuous normal outcomes under the identity link.

CHAPTER FOUR

DATA ANALYSES

4.0 OVERVIEW:

This chapter deals with a summary results of the data analyses of Academic performance on Students Semester Weighted Average (SWA) score and their socio-demographic factor response variables such as gender, entry age, background(Geographical Location of students), as well as the grade level of former school attended. The chapter however seeks to model the data (students' academic performance based on Semester Weighted Average -SWA) using the Generalized Estimating Equation model- GEE (with respect to the four working correlation assumptions specified in methodology.).

4.1 DATA COLLECTION

The consecutive students' Semester Weighted Average (SWA) academic results from (2008-2011) of final year mathematics IV students at 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 formal 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 geographical locations of students were categorized into four zonal belts specifying their respective region of origin. These include the Northern Belt (comprising Northern , Upper East, Upper West Regions) coded as L1, Middle/Central Belt (comprising Ashanti, Brong) was coded as L2, Eastern Belt (Eastern, Volta) 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 GES specification. Grade A schools were coded as 1, grade B schools were coded as 2 and grade C schools were coded as 3. In total, there were 126 students' complete records sampled. Details of the analyses are discussed below.

4.2 EXPLORATORY DATA ANALYSIS

This section illustrates how graphical tools and tables were used to explore the data set. The research results were presented in the form of frequency distribution tables. Computed matrices of Pearson correlation coefficients are shown. Display of graphs showing the merged means of students SWA score with respect to geographical locations and former school attended are shown. This played a contributive factor in description of the observed results using simple percentages (%) and analyses of GEE family of models

4.2.1 The Descriptive Data Analyses of the Study Variables

This sub section of the chapter highlights on the descriptive data analyses for the study with respect to the various study variables. Descriptive statistics are used to describe the basic features of the data gathered from the study. They provide summaries of the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis. The means being the location parameter of the distributions, tell little about the data when left in isolation.

The standard deviation which remains the most common measure of statistical dispersion, measures how widely spread the values in the data set are from the mean. The smaller the standard deviation, the closer are the data points to the mean. The larger the standard deviation, the less representative would be the mean.

Table 4.1: The summary statistics of students SWA scores over the seven semesters

Variable	N	Mean (%)	Std Dev	Minimum (%)	Maximum (%)
SWA1	126	58.95	6.65	24.61	70.50
SWA2	126	57.42	6.67	41.79	77.74
SWA3	126	57.49	8.79	31.81	76.42
SWA4	126	58.80	7.60	37.67	77.84
SWA5	126	56.21	9.75	33.71	79.83
SWA6	126	60.27	7.93	38.33	79.00
SWA7	126	58.13	4.60	45.98	73.66

From table 4.1 above, we observed that the minimum SWA score over all the semesters is in first Semester with a score of 24.61% and the maximum SWA score over all the semesters is in the fifth Semester with a score of 79.83%. It can also be inferred from table 4.1 that the mean SWA of 126 students over the seven semesters were almost similar (i. e. around the average of 60% over the semesters). The small standard deviation depicts how nearer the data points are to the SWA means. The semester with the highest variation of SWA score was semester 5

Table 4.2: Descriptive statistics analyses of the study variables:

Variable	Categories	Frequency	Percentage (%)	Cumu Percent
GENDER	Male	97	77.8	77.8
	Female	28	22.2	100.0
	TOTAL	126	100.0	
GRADED SCHOOL	School A	61	49.2	49.2
	School B	41	32.5	81.7
	School C	23	18.3	100.0
	TOTAL	126	100.0	
GEOGRAPHICAL LOCATION	Location 1 (L1)	15	11.9	11.9
	Location 2 (L2)	47	38.1	50.0
	Location 3 (L3)	33	26.2	76.2
	Location 4 (L4)	30	23.8	100.0
	TOTAL	126	100.0	

We observed from table 4.2 that, out of the 126 students, 97 students were male representing 77.8% and 28 female representing 22.2%.

The graded level of Senior High School students attended, whether Grade A schools, Grade B and Grade C schools are specified in table 4.2. It could be observed that, out of the total 126, 61 students (representing 49.2%) attended a Grade A schools, 41 students (representing 32.5%) attended a Grade B school whilst the 23 remaining students (represents 18.3%) attended a Grade C schools.

Background geographical location of students with respect to whether they come from the Northern Belt coded as L1 (comprising the North, Upper East and Upper

West Regions of Ghana), Middle/Central Belt (comprising Ashanti and Brong) coded as L2, and Eastern Belt (Eastern, Volta) coded as L3 or South/Coastal Belt coded as L4 (comprising Central, Greater Accra, western region) students have been specified. 15 of the respondents representing 11.9% came from the Northern part of Ghana. 47 of the respondents representing 38.1% came from the middle belt. Similarly, 33 (26.2%) students hailed from the Eastern/Volta belt whilst the rest fewer 30 representing 23.8% came from the south-coastal part of Ghana.

The table below shows the Pearson correlation coefficient matrix for the data.

Table 4.3: Pearson correlation coefficients matrix

	SWA1	SWA2	SWA3	SWA4	SWA5	SWA6	SWA7
SWA1	1.0000	0.1931	0.4985	0.2884	-0.0497	-0.0066	0.5179
SWA2	0.1931	1.0000	0.1986	0.2031	0.1465	0.2174	0.5830
SWA3	0.4985	0.1986	1.0000	0.1587	0.1092	0.2681	0.6577
SWA4	0.2884	0.2031	0.1587	1.0000	-0.0054	0.0708	0.4751
SWA5	-0.0497	0.1465	0.1092	-0.0054	1.000	0.3927	0.5314
SWA6	-0.0066	0.2174	0.2681	0.0708	0.3927	1.0000	0.5949
SWA7	0.5179	0.5830	0.6577	0.475	0.5314	0.5949	1.000

It could be observed from table 4.3 that, some pairwise correlation coefficients are weak (<0.5) (e.g. SWA1 versus SWA5, SWA1 versus SWA6) as compared to the others.

It could be seen from the lower triangular matrix of table 4.3 that, SWA7 records stronger correlation with the other semester SWAs as compared to the others, i.e.

0.517, 0.58, 0.66, 0.48, 0.53, 0.59 respectively with SWA1-SWA7 as seen from the last row and column of the Pearson correlation coefficient matrix. There is some evidence of correlation that exist in the data from table 4.3

4.3 THE MEAN STRUCTURE FOR THE STUDY VARIABLES

This section of the chapter discusses the mean structure for students SWA score. The means for all the SWA scores were merged to get an approximate means for students' former school attended and their geographical location of stay.

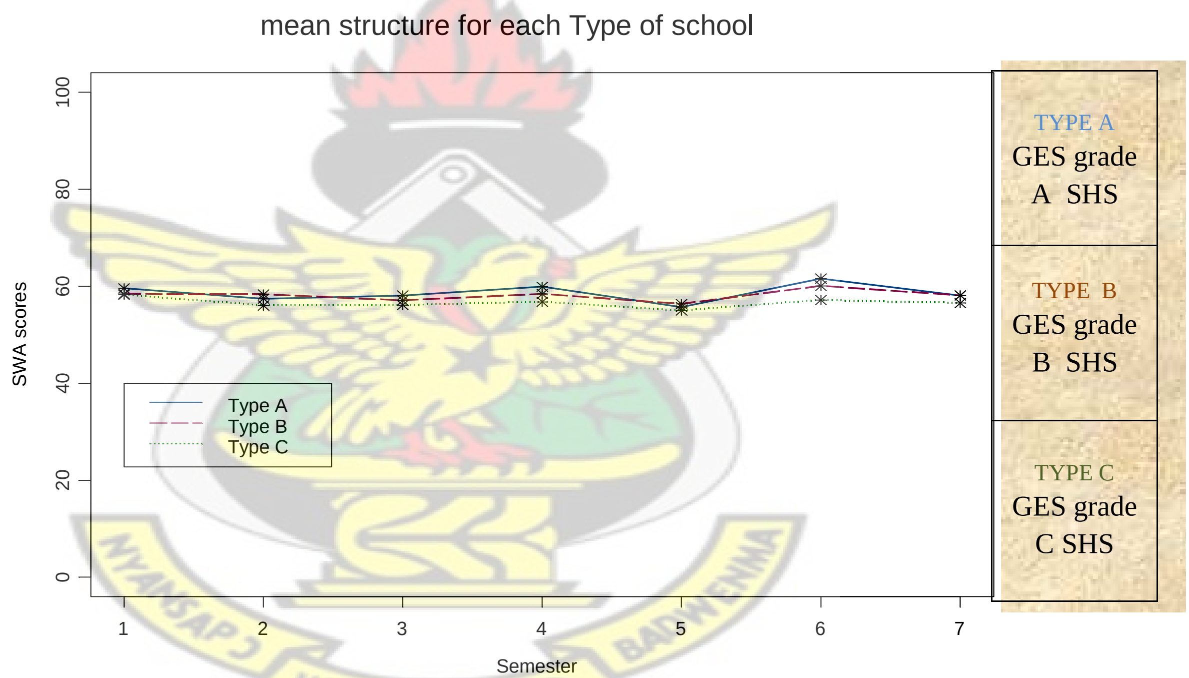

Figure 4.1: ***A Graph of the Mean Structure for Each Type of Graded School.***

From figure 4.1, it can be observed that the mean structure for graded level of former school attended with respect to their SWA's scores is approximately uniform for the three types of schools over semesters (i. e. around the average of 60 over the semesters. See appendix E). The mean structure of SWA scores for type of school seems not to be varying within and between over time (across the seven semesters) on the average. This may indicate that, on the average, academic performance based on SWA score for the type of graded schools were approximately the same over the

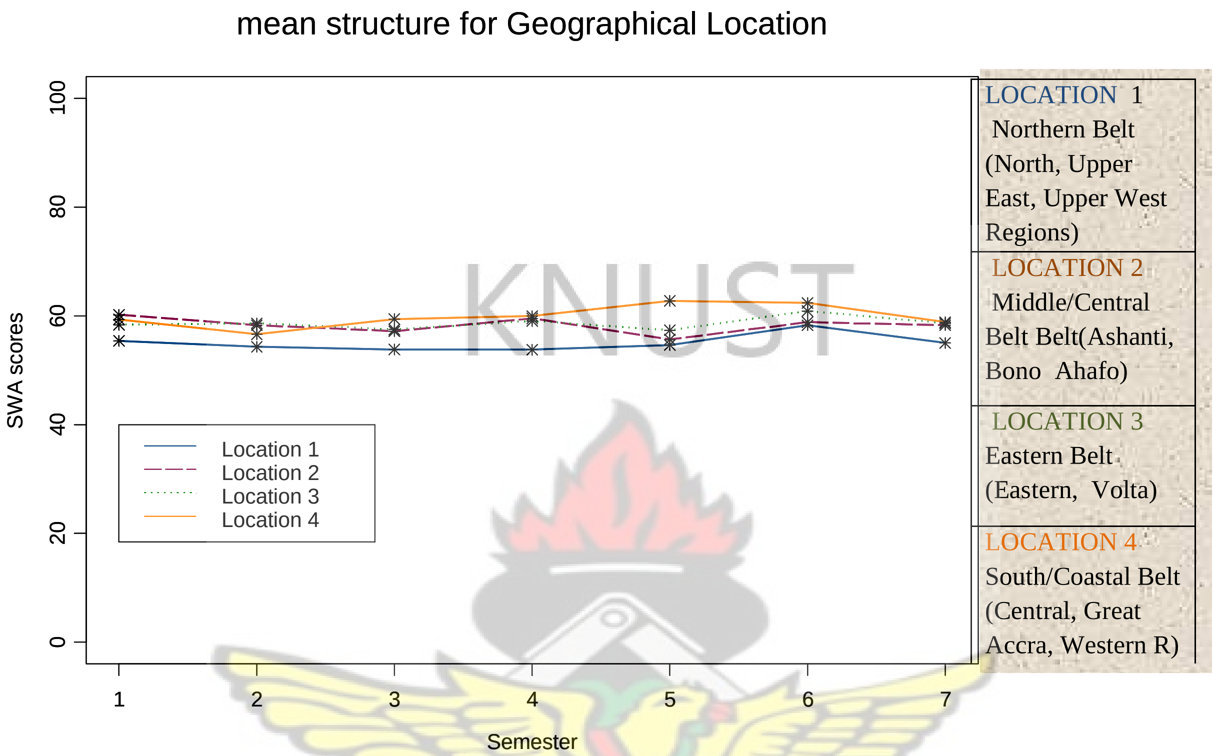

Figure 4.2: The Mean Structure for Geographical Location of stay for Students

From fig 4.2, the SWA means scores for all the Geographical locations for all the seven semesters were variably uniform. Students from Location 4 (L4) seems to have a gradual increase of SWA scores over time (across the seven semesters), L1 also follows in the same way whilst Location 2 seems to have a decrease in SWA scores over time (semesters).

4.4 ANALYSES OF PARAMETER ESTIMATES FOR GEE FAMILY OF MODELS

In Generalized Estimating Equation models, the estimation of the model parameters is paramount and is basically on the working correlation assumptions. In our analyses, this section discusses an analyses of our GEE parameter estimates based on the selected working correlation assumptions. Two sets of linear predictors were fit to these data: a main effect model and a model including condition by linear time interactions with the independent parameters. This are presented in Tables 4.4 and 4.5 respectively. These clearly show the general implicative effect of time on academic performance relative to students' geographical locations and type of graded school attended. The effect may seem to vary by treatment group. Additionally, in preparation for the GEE analysis, the sub sections discusses the working correlation matrix and (ordinary Pearson correlations for these four respective assessments). It is generally advisable to choose a working correlation structure that is similar to the structure of the observed correlations.

Table 4.4: **Estimated Coefficients and Standard Errors for GEE Main Effect Models.**

	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).

Table 4.4 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.

Parameter	INDEPENDENT		EXCHANGEABLE		UNSTRUCTURED		AR(1)	
	Estimate	**Stand Err**	**Estimate**	**Std Err**	**Estimate**	**Std Err**	**Estimate**	**Std Err**
INTERCEPT	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*)
GENDER 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*)
GENDER 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*)

Table 4.5 also summarizes the computations for parameter estimates for our GEE working correlation assumptions with respective time interactions (Independent, exchangeable, unstructured and AR (1) the independent variables with respect to specific time points interaction across the seven semesters. The parameter estimates for the four models are approximately the same for all the assumptions. 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 (Std Err) 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. The abbreviations follow the same way as explained in table 4.4. The proceeding sub-sections discusses the various GEE assumptions as spelt out from the analyses from table 4.4 and 4.5

4.5 GEE MODEL FOR INDEPENDENT

Independent working correlation assumes more than m lag with correlation zero and estimates the parameters within m time points. Independence assumes that there is no correlation within the explanatory variables and the model becomes equivalent to standard normal regression. The "working" correlation matrix is the identity matrix given below.

Working Correlation Matrix for Independent GEE model

$$
\begin{array}{c c}
 & \begin{array}{ccccccc} \text{SWA1} & \text{SWA 2} & \text{SWA 3} & \text{SWA 4} & \text{SWA 5} & \text{SWA 6} & \text{SWA 7} \end{array} \\
\begin{array}{c} \text{SWA 1} \\ \text{SWA 2} \\ \text{SWA 3} \\ \text{SWA 4} \\ \text{SWA 5} \\ \text{SWA 6} \\ \text{SWA 7} \end{array} &
\begin{pmatrix}
1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 \\
0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 \\
0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 \\
0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000
\end{pmatrix}
\end{array}
$$

The matrix is the estimated covariance matrix used for the parameter estimates for the GEE independent model. From our assumption of the GEE independent working correlation displayed in the above matrix, the dependent variables are uncorrelated and are independent of each other across time (for all the seven semesters). Only strong perfectly correlation is realized within and between the dependent variables under study and zeros elsewhere.

4.5.1 GEE Independent Model with Main Effect

Table 4.4 and 4.5 gives the analyses of GEE parameter estimation for the independent model-based and empirical-based standard error estimates. From table 4.4, the analyses of the main effect of GEE parameter estimate for empirical-based (sandwich or robust estimator) and model-based (naive) standard errors are shown.

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 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 4.5 that the parameter estimates for the independent variables (age, gender, geographical locations and type of former school attended) are the same for both empirical and model-based parameter estimates. 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 assumption.

The estimation of the model age parameter (-0.0542) was seen to be statistically insignificant at $\propto = 0.05$ significance level. 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 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.

4.5.2 GEE Independent Model with Linear Time Interactions

In our model including condition by linear time interactions with the independent parameters (see table 4.5), 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 $\alpha = 0.05$ 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).

The standard errors for both model-based and empirical based in the independent GEE with Linear Time Interactions model varies restricting our preference of considering its working correlation assumptions as best fit for the model.

4.6 GEE UNSTRUCTURED (UNSPECIFIED) MODEL

In unstructured working correlation structure specification in GEE modeling, we assume different correlations between any two semester SWA scores for every subject. No constraints are placed on the correlations, which are then estimated from the data. The unspecified or unstructured form would estimate all $n(n - 1) / 2$ correlations of R (n x n identity matrix). Bellow is the working correlation matrix for the unspecified (unstructured) GEE model assumption.

Working Correlation Matrix for the unspecified GEE model assumption.

$$
\begin{array}{c|ccccccc}
 & \text{SWA 1} & \text{SWA 2} & \text{SWA 3} & \text{SWA 4} & \text{SWA 5} & \text{SWA 6} & \text{SWA 7} \\
\hline
\text{SWA 1} & 1.0000 & 0.1795 & 0.4862 & 0.2308 & -0.1101 & -0.0204 & 0.2559 \\
\text{SWA 2} & 0.1795 & 1.0000 & 0.3401 & 0.1745 & 0.2805 & 0.2086 & 0.3443 \\
\text{SWA 3} & 0.4862 & 0.3401 & 1.0000 & 0.1341 & 0.1704 & 0.2694 & 0.4506 \\
\text{SWA 4} & 0.2308 & 0.1745 & 0.1341 & 1.0000 & -0.0629 & 0.0490 & 0.2598 \\
\text{SWA 5} & -0.1101 & 0.2805 & 0.1704 & -0.0629 & 1.0000 & 0.4574 & 0.4224 \\
\text{SWA 6} & -0.0204 & 0.2086 & 0.2694 & 0.0490 & 0.4574 & 1.0000 & 0.3570 \\
\text{SWA 7} & 0.2559 & 0.3443 & 0.4506 & 0.2598 & 0.4224 & 0.3570 & 1.0000
\end{array}
$$

The specification of the working correlation matrix accounts for the form of within subject correlation of responses on the dependent variables. Unstructured R_i working correlation assumes different correlations between any two given semester SWA scores.

4.6.1 GEE Unstructured Main Effect Model

Table 4.4 and 4.5 provides analyses for parameter estimates for GEE unstructured model. It displays the “Analysis of Parameter Estimates”. From table 4.4, the parameter estimates are the same for all the explanatory variables. A critical observation of the standard errors for model-based and empirical-based estimators is marginally different and relatively smaller. 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.

4.6.2 GEE Unstructured Model with Linear Time Interactions

From table 4.5, 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.

4.7 GEE FIRST ORDER AUTOREGRESION (AR-1) MODEL

We now model Academic Performance of students' SWA scores and the independent variables using AR (1) GEE. Autoregressive GEE model weights the correlation within two semester scores by their separated time and hence correlation coefficients diminish for further distances. Similar to exchangeable model, it requires only one estimated parameter. For application of GEE models, one assumes that there are a fixed number of time-points *n* that subjects are measured at.

The matrix bellow is the working correlation matrix for estimating the first order Autoregression GEE model.

Working Correlation Matrix for AR-1 Across the Seven Semesters

	SWA 1	SWA 2	SWA 3	SWA 4	SWA 5	SWA 6	SWA 7
SWA 1	1.0000	0.2214	0.0490	0.0109	0.0024	0.0005	0.0001
SWA 2	0.2214	1.0000	0.2214	0.0490	0.0109	0.0024	0.0005
SWA 3	0.0490	0.2214	1.0000	0.2214	0.0490	0.0109	0.0024
SWA 4	0.0109	0.0490	0.2214	1.0000	0.2214	0.0490	0.0109
SWA 5	0.0024	0.0109	0.0490	0.2214	1.0000	0.2214	0.0490
SWA 6	0.0005	0.0024	0.0109	0.0490	0.2214	1.0000	0.2214
SWA 7	0.0001	0.0005	0.0024	0.0109	0.0490	0.2214	1.0000

Unlike the independent model which uses the identity matrix and the compound symmetry (exchangeable) which merged a constant relationship in its assumptions, first order autoregressive weights the link within two semesters by their estranged time.

4.7.1 GEE AR (1) Main Effect Model

Table 4.4 and 4.5 highlight on the analyses of parameter estimates and their respective standard errors for both model based and empirical based estimation. We can observe from table 4.4 that the results are not different from the previous analyses for independents, exchangeable and Unstructured GEE models.

The parameter estimate for all the variables are the same for empirical and model-based. However, differences exist in the variations of the standard errors for empirical and model-based.

The intercept has a significantly higher standard error estimate of 5.1466 and 4.3116 for naïve and robust respectively. Gender status of students 0 (representing male) records the highest Standard Error Estimate (SEE) of 4.0504 and 4.9757 and for gender 1 (representing females) has SEE of 4.1843 and 5.1466 respectively. The others are spread between 0.8018 to 0.9769 for the naïve and 0.2135 to 1.3203 for empirical based.

A critical observation of the standard errors for model-based and empirical-based estimators is marginally different and relatively smaller. It could be seen that the variations of these standard error estimates reduce the effectiveness of considering the GEE first order Autoregression working correlation assumption as not well fitted for the model.

The contrasts within Gender statuses for both male and female students as well as geographical location 1 and 4 continue to be statistically significant in the model.

4.7.2 GEE AR (1) Model with Linear Time Interactions

The results of the analyses from table 4.5 is relatively similar to almost all the GEE family of models used for this study. The AR-1 records the highest parameter estimate of 60.420 for the intercept at a highly significant level. The interaction of the gender effect in the model is not significant. 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 $\alpha = 0.05$ with variations in model-based and empirical standard error estimate

4.8 GEE EXCHANGEABLE (COMPOUND SYMMETRY) MODEL

The exchangeable working correlation specification allows for constant correlations between any two (2) measurements within a subject for all the time points. As such, only one parameter needs to be estimated. The types of working correlation examined in this model in order to measure the relationship between the student's SWA scores over time across the seven semesters is given bellow. The specification of the working correlation matrix in estimating the covariance of the parameter estimates is based on the four assumptions such that, the working correlation specification allows for constant correlations between any two (2) measurements within a subject i for all the time points across the seven semesters. The matrix below is the Working Correlation Matrix for Exchangeable GEE Model.

Working Correlation Matrix for Exchangeable GEE Model

$$
\begin{array}{c c}
 & \begin{array}{ccccccc} \text{SWA 1} & \text{SWA 2} & \text{SWA 3} & \text{SWA 4} & \text{SWA 5} & \text{SWA 6} & \text{SWA 7} \end{array} \\
\begin{array}{c} \text{SWA 1} \\ \text{SWA 2} \\ \text{SWA 3} \\ \text{SWA 4} \\ \text{SWA 5} \\ \text{SWA 6} \\ \text{SWA 7} \end{array} &
\begin{pmatrix}
1.0000 & 0.2180 & 0.2180 & 0.2180 & 0.2180 & 0.2180 & 0.2180 \\
0.2180 & 1.0000 & 0.2180 & 0.2180 & 0.2180 & 0.2180 & 0.2180 \\
0.2180 & 0.2180 & 1.0000 & 0.2180 & 0.2180 & 0.2180 & 0.2180 \\
0.2180 & 0.2180 & 0.2180 & 1.0000 & 0.2180 & 0.2180 & 0.2180 \\
0.2180 & 0.2180 & 0.2180 & 0.2180 & 1.0000 & 0.2180 & 0.2180 \\
0.2180 & 0.2180 & 0.2180 & 0.2180 & 0.2180 & 1.0000 & 0.2180 \\
0.2180 & 0.2180 & 0.2180 & 0.2180 & 0.2180 & 0.2180 & 1.0000
\end{pmatrix}
\end{array}
$$

The interaction effect between rows and columns variables of the exchangeable working correlation matrix is constant across all the seven time points and is estimated to be 0.2180.

4.8.1 GEE Exchangeable Main Effect Model

Information about the GEE Model (exchangeable) is displayed in table 4.4 and 4.5 for both empirical standard error and model-based standard error estimates respectively. The results of fitting the model are however shown.

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. No significant gender effect was found. The age parameter estimates are statistically not significant at $\propto = 0.05$. for both model.

In the parameter estimate for location, only the difference between location 1 and location 4 (-3.6729) was significant with standard error of 1.338. 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 contrasts for type of former school attended A and B were also insignificant and have estimation of standard error different in the respective empirical and model based parameter estimates. The standard errors for empirical and model based as seen from table 4.5 are relatively close to each other and hence fit the model estimation relatively well. We notice that if

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

Then the two 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.

4.8.2 GEE Exchangeable Model with Linear Time Interactions

In linear time interactions with the exchangeable parameters as seen table 4.5, the parameter estimates were approximately the same. From this model, only the intercept, location 1 and time interaction with geographical location 2 parameter estimates were found to be statistically 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 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 others.

4.9 GENERALIZED WALD TESTS FOR MODEL COMPARISON: CONTRAST RESULTS FOR GEE ANALYSIS

In order to interpret the group-related effects of the GEE models, it would be helpful to compare these models statistically to determine if the groups by time interaction terms (i.e. students geographical locations and graded level of former school types in correspondence to their SWA scores) are jointly significant or not.

Because GEE model parameters are estimated using quasi-likelihood procedures, there is no associated likelihood underlying the model. Thus, the usual likelihood ratio tests cannot be applied to compare the above models. To compare the above GEE models, however, we can construct a multi-parameter Wald test to test the joint null hypothesis that a set of β's 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.

The table below outlines the contrast result for GEE compound symmetry (Exchangeable) analyses.

Table 4.6: Contrast Result for GEE Analysis for main effect model

Contrast	DF	Chi-Square	Pr > Chi Sq
LOC 1 vs LOC 2	1	4.63	0.0314
LOC 1 vs LOC 3	1	3.66	0.0526
LOC 2 vs LOC 3	1	0.01	0.9215
SCH A vs SCH B	1	0.81	0.3673

From table 4.6, a test of the contrasts of each of the parameter estimate (with respect to four locations and three types of graded former school type attended) under chi-square distribution as confirmed by their respective p-values displayed.

In the present study, comparing the models contrast for the locations, we get the computed chi-square value for the contrast of location 1 and 2 to be 4.63 which yields p-value of Pr > Chi Sq $p = 0.0314$ and hence significant. In the same way the contrast of location 1 and 3 is marginally significant with a computed chi-square of 3.66 and $p = 0.0526$. There is no statistical significance of the contrast between location 2 and 3 as it records $x^2 = 0.01$ with a higher $p - value$ of 0.9215. Additionally, none of the individual tests of the type of former school attended in Table 4.6 are significant at $\alpha = 0.05$ either.

Thus, the GEE exchangeable model is preferred to all the other models.

4.10: HYPOTHESIS TESTING

In consideration to an output revealed from SAS (see from table 4.4-4.6), we test the appropriateness or otherwise of each of the model parameters such that one of the coefficient of the $\beta_i \neq 0 \quad for\ at\ least\ one\ i$. The appropriate hypothesis is given as

$$H_0: \beta_1 = \beta_2 = \beta_3 \ldots = \beta_k = 0 \ against\ the\ alternative\ that$$

$$H_1: \beta_i \neq 0 \quad for\ at\ least\ one\ i$$

At $\alpha = 0.05$ *level of significance.*

Where β_i are the model parameters (independent variables). A test of each of the parameter estimate under chi-square distribution from table 4.6 reveals that, the contrast for students' geographical locations is marginally significant at $\alpha = 0.05$ with P-values estimated to be less than α = 0.05. However, gender, age and graded level of former school type students attended was contrasted to be statistically insignificant.

Conclusion: Hence the Model is generally significant at 95% (0.05) significant level since

$$\beta_i \neq 0 \quad for\ at\ least\ one\ i$$

(This means that, at least one of the parameters is significant at $\alpha = 0.05$ significance level).

4.11 THE BEST FITTED WORKING CORRELATION ASSUMPTION FROM GEE FAMILY OF MODELS

It could be inferred from the SAS output that, the statistical output with respect to the GEE independent, exchangeable and unstructured model for both empirical and model-based standard error estimates the results of fitting the model parameter are identical, but the standard errors for the various GEE model varies within and across the parameters. However, 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. The coefficient of the interaction term $\beta's$ is marginally significant with the exchangeable model.

It is generally advisable to choose a working correlation structure that is similar to the structure of the observed correlations. This is because, although the GEE is robust to misspecification of the correlation structure, efficiency is increased to the extent that the specified structure is correct. In the present case, the unspecified structure does not appear like a good choice since the correlations are not approximately equal. Also, neither the AR (1) nor the independent structures appear reasonable because the correlations within the time lag over the seven semesters vary as revealed from the variations in their standard error estimations.

In consideration to the GEE model fit for main effect and linear time interaction, we opted for the main effect model since the linear time interaction proved all the parameter estimates to be statistically insignificant at 5% level of significance. Thus, an exchangeable working correlation structure for GEE main effect model appears to be the most reasonable choice for these data. The table bellow (table 4.7) shows the summary analyses of the GEE Exchangeable working correlation assumption for our analyses. The table gives a summary report for GEE Exchangeable main effect model with specified model-based and empirical standard errors, 95% confidence intervals and comparative $P - values$ with standard normal references.

Table 4.7: Parameter Estimates of Model-Based and Empirical-based Standard Errors for GEE Exchangeable Model

Analysis Of GEE Parameter Estimates							
Model-Based Standard Error Estimates for GEE Exchangeable Model							
Parameter		**Estimate**	**Std Err**	**95% Confidence Limits**		**Z**	**Pr > |Z|**
INTERCEPT		59.9731	5.2739	49.6364	70.3099	11.37	<.0001
AGE		-0.0542	0.2132	-0.4720	0.3636	-0.25	0.7992
GENDER 0	0	-1.3694	0.9589	-3.2488	0.5100	-1.43	0.1533
GENDER 1	1	0.0000	0.0000	0.0000	0.0000	.	.
LOC 1	1	-3.6729	1.5878	-6.7849	-0.5609	-2.31	0.0207
LOC 2	2	-0.7283	1.0106	-2.7091	1.2524	-0.72	0.0411
LOC 3	3	-0.6176	1.1015	-2.7766	1.5413	-0.56	0.5750
LOC 4	4	0.0000	0.0000	0.0000	0.0000	.	.
SCH *A*	A	1.1160	1.1713	-1.1797	3.4118	0.95	0.3407
SCH *B*	B	0.3551	1.2313	-2.0582	2.7685	0.29	0.7730
SCH *C*	C	0.0000	0.0000	0.0000	0.0000	.	.
Working corr		0.2180	.	.	.	.	.
Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates for GEE Exchangeable Model							
Parameter		**Estimate**	**Std Err**	**95% Confidence Limits**		**Z**	**Pr > |Z|**
INTERCEPT		59.9731	5.2739	49.6364	70.1764	11.52	<.0001
AGE		-0.0542	0.2154	-0.4765	0.3680	-0.25	0.8012
GENDER 0	0	-1.3694	0.9516	-3.2345	0.4957	-1.44	0.1501
GENDER 1	1	0.0000	0.0000	0.0000	0.0000	.	.
LOC 1	1	-3.6729	1.3383	-6.2960	-1.0498	-2.74	0.0061
LOC 2	2	-0.7283	0.8666	-2.4267	0.9701	-0.84	0.0406
LOC 3	3	-0.6176	1.1612	-2.8935	1.6583	-0.53	0.5948
LOC 4	4	0.0000	0.0000	0.0000	0.0000	.	.
SCH *A*	A	1.1160	1.1420	-1.1224	3.3544	0.98	0.3285
SCH *B*	B	0.3551	1.2027	-2.0021	2.7123	0.30	0.7678
SCH *C*	C	0.0000	0.0000	0.0000	0.0000	.	.
Work. corr		0.2180					

Table 4.7 gives a snapshot of the summary statistics of GEE Exchangeable working correlation assumption modeled and downloaded directly from SAS version 9.1 outputs. The main effect model from the table gives only a statistical significance to students' geographical locations at $\alpha = 0.05$ significance level as confirmed from the contrast effects analysed in (table 4.6). no significant effect was seen for students entry age, gender as well as former school attended.

4.12 CHAPTER SUMMARY:

The chapter sought to analyze the data collected from students AP in their SWA scores and their associated socio-demographic factors. This chapter has described the various GEE approach with respect to its four underlying assumptions for such a longitudinal data analysis. The analyses focused on the GEE main effect and linear time interaction model for Independence; exchangeable, AR-1 and unstructured (each with their empirical and model-based standard error estimates) models. Generalized Estimating Equations (GEE) provides a practical method with good statistical properties to model data that exhibit association but cannot be modeled as multivariate normal. This approach has several features which makes it particularly useful and popular. Because it is a generalization of GLM, many types of dependent variables can be accommodated within the GEE family of models. Statistical software packages used to perform the GEE analysis was SAS 9.1 system. GEE provides regression estimates that are "population-averaged" rather than the "subject specific" estimates of the mixed-effects regression models.

CHAPTER FIVE

SUMMARY, CONCLUSSIONS AND RECOMMENDATIONS

5.0 OVERVIEW:

The chapter gives a brief report on the summary of the study, findings and conclusion from the statistical analyses conveyed in the study as well as recommendations generated from the findings.

5.1 SUMMARY:

The study sought to investigate into the socio-demographic factors of students such as age gender, geographical locations and graded level of students' former school attended as independent variables in relation to their Academic performance of SWA scores in KNUST mathematics department. Below are the summary reports from the research carried out.

The study began with introduction in correspondence to the problem statements for the study. This study attempts to find the factors that determine the level of achievement of the students SWA scores in the KNUST Mathematics Department in terms of students' age, gender, graded level of students' former SHS attended and the geographical location of students. The study however sought to find out whether these socio-demographic factors also affect students Academic Performance. Review of related literature was analyzed for the study highlighting on what people have found on the issue of academic performance and variables accounting for its reality or otherwise.

The methodology used for the study was the Generalized Estimating Equation (GEE) family of models. The researcher chose this because, Generalized Estimating Equations (GEE) extend the GLM algorithm to accommodate correlated data such as these. The algorithms of Generalized Estimating Equations (GEE) are based on (Liang & Zeger, 1986) and (Diggle & Heagerty, 2002). This helped in analyzing the existing correlation between the study variables and defining a best fitted model for the study.

Data collection was gathered from students SWA score across time (for the seven covered semesters) spelling out in details the academic performance average score of every student in the department. These students censusly sampled were given sample questionnaire to respond to certain factors leading to the researcher enlisted independent variables. The data collected were analyzed using SAS 9.1 version. Other supportive statistical software package used for the analyses were SPSS 16, MINITAB 14 and the Microsoft Excel 2007 for coding and further supportive analyses for the project report. One hundred and twenty-six (126) students were sampled from KNUST Mathematics Department.

The analyses reveal that, out of the four independent variables studied, students' geographical location was marginally significant and very predictive of the academic performance of students as compared to the other entry variables.

However, the general analyses was considered and tested to be significant under the assumptions defined by the model-based and empirical-based standard error estimate for the GEE exchangeable chosen as the best fit model. The generalized Wald test for model comparison for contrast results for the GEE analyses was conducted to confirm its suitability. 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. However, we note the proximity of the standard errors for both empirical and model based in the GEE model for Exchangeable confirming its suitability for the actual model.

Below are the Major findings:

Students' entry age, gender as well as their graded level of former school attended does not necessarily affect academic performance in their SWA scores since their parameter estimation and contrast effects were found to be statistically insignificant, and hence play no active important role in ensuring the academic performance achievements of student SWA scores in the KNUST Mathematics department.

Coefficient Estimation of the Study Parameters and the contrast interaction effects in the various GEE models for main effect and time interaction effects reveal that, students' geographical locations were significant in as far as achievements of SWA scores in the KNUST Mathematics department is concern.

The estimated coefficients of the four models were also very close and identical. This reaffirmed the closeness of the measures for comparison and the consistently supported claim of (Zeger and Liang, 1986) that such results are expected when working correlations are misspecified. In general, no significant age effect was found.

5.2 CONCLUSIONS

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).

5.3 RECOMMENDATIONS:

After careful analyses of the study, we recommend that;

- Mathematics Education should be strengthened in the Northern, Upper East and Upper West regions since the contrast effect of their average performance with other locations in Mathematics decreases over time.
- Further research should be conducted to check individual differences existing among the students that may account for differences in Academic performance in Institutions, (For example using the Random Effect Model)

BIBLIOGRAPHY

Allen, W. (1992). *The color of success: African-American college student outcomes*. American Educational Research Journal, 14(3), 317-330.

Amadeahe, F. K. (2002). *Fundamentals of Educational Research Methods.* Mimeograph, UCC, Cape Coast.

Alnerst, S. F. (1991). *Motivation factors Influencing Students who attain Valediction or Salutatorian Statuses in University of Arizona, (*PHD *1991).* International Dissertation abstract 51 (12).

Bell, L.R. (2010). Update on Academic Performance; affective variables in learning mathematics. *Review of Educational Research, 46, 293-311.*

Birch, E., & Miller, P. (2004). *The determinants of students' tertiary academic success, in, Conference proceedings*. Quantitative tools for microeconomic policy analysis, Canberra.

Brewer, C. (2008). *Using generalized estimating equations with Regression splines to improve analysis of butterfly Transect data*. Mphil Thesis, University of Andrews.

Burstein, L. (1980). *The analysis of multilevel data in educational research and evaluation.* In Berliner, D. editor, Review of Research in Education, volume 8, pages 158-233. American Educational Research Association, Washington, DC, 1980.

Burton, P., Gurrin, L., & Sly, P. (1998). *Tutorial in biostatistics: extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level modeling.* Statistics in Medicine, 17: 1261-1291, 1998.

Craft, M. (1970). *Family Class and Education-Agenda: London*. Longman Group Ltd.

Caberra, K. A., & Nora, P. (1996). *High School Vocational Education, Apprenticeship, and Earnings: A Comparison of Germany and the United States.* Vierteljahreshefte zur Wirtschaftsforschung, Heft 1, 1994, pp. 10 - 18.

Eshun F. (1999). *The Pattern of Mathematics Achievement of Senior Secondary Schools in Ghana*. A Journal of Science and Mathematics Education, UCC.

Evans, M. (1999). *School-leavers' transition to tertiary study: a literature review*.

Working *Index*, 56, 293-121.

Fitzmaurice, G.M. (1995). *A caveat concerning independence estimating equations with multivariate binary data*. Biometrics, 51:309-317, 1995.

Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd edition. Halstead Press, New York,

Gordon, J., & Maura, S. (2009). *Applications of GEE Methodology Using the SAS System.* SAS Institute Inc., Cary, NC.

Gonzale, P. G. (1994). *Application of Statistical Models in educational issues*. 2nd edition. Halstead Press, New York.

Gur, P., Frazer, R., & Beam, T. G. (1977). *Factors affecting Students' Academic Performance.* University of Sharon Larisa Segrest of South Florida St. Petersburg

Goldman, J. M. (1981). *Time-Series Analysis: A Comprehensive Introduction for Social Scientists.* Cambridge University Press, New York.

Goethe, E., & Fiske, P. (2001). *The uneven playing field of school choice: evidence from New* Zealand, Journal of Policy Analysis and Management, 20 (1): p 43.

Ghana Edcation Service[GES], (2006). *The computer School Selection and Placement System(CSSPS),* www.modernghana.com/.../computer-selection-and-placement-system-reviewed.html. Date access, 2nd February, 2011.

Hedeker D. (1989). *Random Regression Models with Autocorrelated Errors*. Ph.D. thesis, University of Chicago, Department of Psychology, 1989.

Hosmer, D. W., & Lemeshow, S. (2001). Applied logistic regression, 2nd edition, Wiley, New.

Hoxby, J. (2000). *Multilevel Analysis: Techniques and Applications*. Erlbaum, Mahwah, NJ, 2002.

Kerlinger, F. N. (1973). Foundations of Behavioural Research. New York: Holt, Rinehart and Winston Inc.

Kiamanesh, A. R. (2002). *Factors affecting Iranian students' achievement in Mathematics*. Teacher Training University, Iran.

Kenteyky Adult Education [KYAE], (2009). *National Center for Education Statistics*, "Projections Of Education Statistics to 2012" 101-2009.

Kiang, T. T., Trivina, K., & Hogan, D. (2009). *Using GEE to Model Student's*

Satisfaction: A SAS ® Macro Approach, Centre for Research in Pedagogy and Practice, Nanyang Technological University, Singapore Paper 251-2009 Practice, Nanyang Technological University, Singapore Paper 251-2009.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models*. 2nd edition, Chapman and Hall, New York, 1989.

Nelder, J. A., & Wedderhrn, R. W. M. (1972). *Generalized linear models*. Journal of the Royal Statistical Society, Series A **,** 135:370-384,1972.

Nora, A., & Cabrera, S. (1996). *The role of perceptions of prejudice and Discrimination on the Academic Performance of Students.* Abstract 45(1/2) Tanti Press Ltd.

Papanastasiou, C. (2002). *School, teaching and family influence on student attitudes toward science:* Based on TIMSS data Cyprus. Studies in Educational Evaluation, 28, 71-86.

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

Schultte, J. (1998). *Discovering an optimal property of the Median"*, Journal citation: Mathematics teacher; v. 92, n. 8, pp. 692-703.

Scott, D., & Waltert Y. (2005). *Retention, completion and progression in tertiary education 2003, Wellington*: The Educational Forum 4(2): 460-483.

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

Stodalsky, S.S., Salk, S. & Glaessner, B. (1991). *Student views about learning math and social Stratification and Mobility*. 22: 187–216 Studies in Educational Evaluation, 26, 27-42.

Soyibo, K. & Thiessen, V. (2007). *The impact of factors on trajectories that lead to Non-completion of high school and lack of post-secondary education among those with high reading competencies at age 15.* Quebec: Human Resources and Social Development Canada.

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

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

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

Wolfinger, R. D. (1993). *Covariance structure selection in general mixed models. Communications in Statistics, Simulation and Computation*. 22: 1079-1 106,1993. York, 2000.

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

Zheng, B. (2000). *Summarizing the goodness of fit on generalized linear models for longitudinal data.* Statistics inc.

Zajonca, C. (1976). *Generalized Estimating Equation Models for Correlated Data: A Review with Applications.* American Journal of Political Science, 45, 470–490.

APPENDIX A

KWAME NKRIMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF MATHEMATICS
QUESTIONNAIRE TO STUDENTS

STUDENT'S INDEX NO:…………………………………………………………….

Overview: The purpose of this questionnaire is to collect information about the variables that affect the Academic Performance (A.P.) MATH 4 students in KNUST with respect to your socio-demographic data (age, gender), formal school attended as well as your geographical location as part of my MPHIL thesis at the above University. Your responses to these questions will be treated and kept strictly confidential since it is for academic purpose only.

Please tick [√] or underline the response option that indicates your view or opinion on each of the given statements00000000.00

Section A (Sex/Gender Status Of Students Versus A.P.)

1. SEX/GENDER: Male [ ] Female [ ]

Section B (Age Distribution Of Students)

2. Please indicate your present age……………………………………………

Section C (Geographical Location of Students)

3. Which part of Ghana (in terms of the following tabled regions) do you come from?

REGION	TICK	REGION	TICK
i Northern	[ ]	vi Ashante	[ ]
ii Upper East	[ ]	vii Eastern	[ ]
iii Upper West	[ ]	viii Volta	[ ]
iv Brong Ahafo	[ ]	ix Central	[ ]
v Western	[ ]	x Greater Accra	[ ]

4. Are you a foreign student? A. NO [ ] B. YES [ ]

5. If yes state which country you come from…………………………………..

Section D (Graded Level Of Formal School Attended)

6. What is the name of your formal school?..

7. Which region in Ghana is your formal school located?..................................

THANK YOU FOR YOUR COOPERATION

APPENDIX B

SAS CODE FOR GEE MODEL ON ACADEMIC PERFORMANCE

```
libname iod 'C:\swa';

/*** Importation of the data***/

PROC IMPORT OUT= iod.swa
            DATAFILE= "C:\iod.txt"

         DBMS=TAB REPLACE;

     GETNAMES=YES;
   DATAROW=2;

RUN;

 proc print data=iod.swa ; run;

/*** Manupulation of the data ***/

 data iod.nswa;
 set iod.swa;
 if SEX=1 then gender=1;
 if SEX=2 then gender=0;
  if school='A' then do ;ta=1;end; else ta=0;

  if school='B' then do; tb=1;end; else tb=0;
  if school='C' then do; tc=1;end; else tc=0;

  drop school SEX;
 run;
 proc print;run;

proc sort data = iod.nswa;
by SN;
run;

/*****Means, Covariances, and Correlations*******/

proc freq data=iod.nswa;
run;
proc corr data = iod.nswa cov;
var swa1- - swa7;
run;

/*** Changing the multivariate data to univariate data***/

 data iod.swa;
 set iod.nswa;
 swa=swa1;tme=1;output;
 swa=swa2;tme=2;output;
 swa=swa3;tme=3;output;
 swa=swa4;tme=4;output;
 swa=swa5;tme=5;output;
 swa=swa6;tme=6;output;
 swa=swa7;tme=7;output;
  drop swa1-swa7;
 run;
```

```
proc print data=iod.swa;run;

data iod.newswa;
set iod.newswa ;
tme2=tme*tme;
run;

/*** Splitting dataset by the three types of schools ***/

proc sort data=iod.newswa;
by form_sch_ tme;
run;

data ty1;
set iod.newswa;
where form_sch_=1;
run;

data ty2;
set iod.newswa;
where form_sch_=2;
run;
data ty3;
set iod.newswa;
where form_sch_=3;
run;

proc means data=ty1;
by tme;
var swa;
output out=m1 mean=meant1;
run;

proc means data=ty2;
by tme;
var swa;
output out=m2 mean=meant2;
run;

proc means data=ty3;
by tme;
var swa;
output out=m3 mean=meant3;
run;

/*** Merging the means of the types of schools***/
data typemean;
set m1 m2 m3;
merge m1 m2 m3;
by tme;
keep tme meant1 meant2 meant3;
run;

proc print data=typemean;
run;

/*exploration of the evolution mean in each type*/
```

```
goptions reset=all ftext=swiss device=psepsf gsfmode=replace
rotate=landscape i=join;
proc gplot data=typemean;
plot meant1*tme=1 meant2*tme=2 meant3*tme=3 /overlay haxis=axis1
vaxis=axis2;
symbol1 c=red v=star i=std1mjt w=2 mode=include;
symbol2 c=black v=dot i=std1mjt w=2 mode=include;
symbol3 c=blue v=triangle i=std1mjt w=2 mode=include;
axis1 label=(h=2 'Semesters') value=(h=1)
      order=(1 to 6 by 1) minor=none;
axis2 label=(h=2 A=90 'SWA scores') value=(h=1)
      order=(30 to 90 by 3) minor=none;
title h=2 'Average evolution for each type of school means';
run;

goption reset=all i=join;
proc gplot data=typemean;
plot meant1*tme=1 meant2*tme=2 meant3*tme=3/overlay haxis=axis1
vaxis=axis2 ;
symbol1 c=red v=star i=join h=1 w=1 mode=include;
symbol2 c=black v=dot i=join h=1 w=1 mode=include;
symbol3 c=blue v=triangle i=std1mjt w=1 mode=include;
axis1 label=(h=1 'Semesters') value=(h=1.5) minor=none order=(1 to 6
by 1);
axis2 label=(h=1 A=90 'SWA scores') value=(h=1.5);
run;
quit;

 /* exploration of the evolution of the overall mean*/
goptions reset=all ftext=swiss device=psepsf gsfmode=replace
rotate=landscape i=join;
proc gplot data=ama.nwswa;
plot respons*time / haxis=axis1 vaxis=axis2;
symbol c=red i=std1mjt w=2 mode=include;
axis1 label=(h=2 'Weeks') value=(h=1)
      order=(0 to 52 by 4) minor=none;
axis2 label=(h=2 A=90 'Visual') value=(h=1)
      order=(40 to 60 by 5) minor=none;
title h=2 'Average evolution,
      with standard errors of means';
run;
quit;

/*** Splitting dataset by the four locations ***/
proc sort data=iod.newswa;
by location tme;
run;

data loc1;
set iod.newswa;
where location=1;
run;

data loc2;
set iod.newswa;
where location=2;
run;

data loc3;
```

```
set iod.newswa;
where location=3;
run;
data loc4;
set iod.newswa;
where location=4;
run;
proc means data=loc1;
by tme;
var swa;
output out=l1 mean=lmeant1;
run;

proc means data=loc2;
by tme;
var swa;
output out=l2 mean=lmeant2;
run;

proc means data=loc3;
by tme;
var swa;
output out=l3 mean=lmeant3;
run;
proc means data=loc4;
by tme;
var swa;
output out=l4 mean=lmeant4;
run;

/*** Merging the means of the locations***/
data typemean1;
set l1 l2 l3 l4;
merge l1 l2 l3 l4;
by tme;
keep tme lmeant1 lmeant2 lmeant3 lmeant4;
run;

proc print data=typemean1;
run;
/*** modelling the GEE***/
/***GEE MODEL FOR INDEPENDENCE***/
 proc genmod data=iod.swa;
class sn gender location form_sch_;
model swa = age gender location form_sch_ / dist   = normal
link   = identity covb;
repeated subject=sn / corrw type=ind modelse;
run;
proc genmod data=iod.swa;
class sn gender location form_sch_;
model swa = age gender location form_sch_ location*tme form_sch_*tme
/ dist   = normal
link   = identity covb;
repeated subject=sn / corrw type=ind modelse;
run;
/***GEE MODEL FOR EXCHANGEABLE***/
proc genmod data=iod.swa;
class sn gender location form_sch_;
model swa = age gender location form_sch_ / dist   = normal
```

```
link    = identity covb;
repeated subject=sn / corrw type=exch modelse;
run;
proc genmod data=iod.swa;
class sn gender location form_sch_;
model swa = age gender location form_sch_ location*tme form_sch_*tme
/ dist    = normal
link    = identity covb;
repeated subject=sn / corrw type=exch modelse;
run;

/***GEE MODEL FOR UNSTRUCTURED***/
proc genmod data=iod.swa;
class sn gender location form_sch_;
model swa = age gender location form_sch_ / dist    = normal
link    = identity covb;
repeated subject=sn / corrw type=un modelse;
run;
proc genmod data=iod.swa;
class sn gender location form_sch_;
model swa = age gender location form_sch_ location*tme form_sch_*tme
/ dist    = normal
link    = identity covb;
repeated subject=sn / corrw type=un modelse;
run;

/***GEE MODEL FOR AR(1)***/
proc genmod data=iod.swa;
class sn gender location form_sch_;
model swa = age gender location form_sch_ / dist    = normal
link    = identity covb;
repeated subject=sn / corrw type=AR(1) modelse;
run;
proc genmod data=iod.swa;
class sn gender location form_sch_;
model swa = age gender location form_sch_ location*tme form_sch_*tme
/ dist    = normal
link    = identity covb;
repeated subject=sn / corrw type=AR(1) modelse;
run;


/***GEE MODEL FOR EXCHANGEABLE as best fit WITH CONTRASTS***/

proc genmod data=iod.swa;
class sn gender location form_sch_;
model swa = age gender location form_sch_ / dist    = normal
/solution
link    = identity covb;
repeated subject=sn / corrw type=exch modelse;
contrast 'location 1 vs location 2' location 1 -1 0 0;
contrast 'location 1 vs location 3' location 1 0 -1 0;
contrast 'location 2 vs location 3' location 0 1 -1 0;
contrast '  form_sch_ 1 vs  form_sch_ 2' form_sch_ 1 -1 0;
contrast 'tme*location 1 vs tme*location 2' location 1 -1 0 0;
contrast 'tme*location 1 vs tme*location 3' location 1 0 -1 0;
contrast 'tme*location 2 vs tme*location 3' location 0 1 -1 0;
contrast '  tme*form_sch_ 1 vs  tme*form_sch_ 2' form_sch_ 1 -1 0;
run;
```

APPENDIX C

DESCRIPTIVE STATISTICS TABLES FOR THE DATA

Table 4.2: The descriptive statistics analyses of Student's SWA(s) Achievement

	N	Range	Minimum	Maximum	Sum	Mean		Std. Dev	Var
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Erro	Statistic	Statistic
SWA1	126	45.89	24.61	70.50	7439.18	59.0411	.58902	6.61170	43.715
SWA2	126	77.74	.00	77.74	7185.52	57.0279	.74938	8.41180	70.758
SWA3	126	44.61	31.81	76.42	7252.62	57.5605	.78427	8.80345	77.501
SWA4	126	40.17	37.67	77.84	7417.54	58.8694	.67729	7.60261	57.800
SWA5	126	46.12	33.71	79.83	7066.43	56.0828	.86145	9.66979	93.505
SWA6	126	40.67	38.33	79.00	7581.61	60.1715	.70168	7.87631	62.036
SWA7	126	27.68	45.98	73.66	7323.83	58.1256	.41189	4.62345	21.376
Valid N (listwise)	126								

FREQUENCY TABLE FOR THE STUDY VARIABLES

Table 4.5: The Age Distribution for sampled Students

		Frequency	Percent	Valid Percent	Cum Percent
AGE	21	4	3.2	3.2	3.2
	22	32	25.4	25.4	28.6
	23	41	32.5	32.5	61.1
	24	22	17.5	17.5	78.6
	25	16	12.7	12.7	91.3
	26	2	1.6	1.6	92.9
	27	3	2.4	2.4	95.2
	28	2	1.6	1.6	96.8
	30	2	1.6	1.6	98.4
	34	1	.8	.8	99.2
	35	1	.8	.8	100.0
	Total	126	100.0	100.0	

Table 4.7: Frequency Distribution Table Showing The Gender Of Sampled Students

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	MALE	98	77.8	77.8	77.8
	FEMALE	28	22.2	22.2	100.0
	Total	126	100.0	100.0	

Table 4.6: Frequency Distribution For Geographical Location Of Students

		Frequency	Percent	Valid Percent	Cumulative Percent
Region	Northern Belt(North, Upper East, Upper West Regions)	15	11.9	11.9	11.9
	Middle Belt/Central Belt(Ashanti, Brong)	48	38.1	38.1	50.0
	Eastern Belt(Eastern, Volta)	33	26.2	26.2	76.2
	South/Coastal Belt(Central, Greater Accra)	30	23.8	23.8	100.0
	Total	126	100.0	100.0	

Table 4.8: A Frequency Distribution for Graded School Status of Students

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	62	49.2	49.2	49.2
	B	41	32.5	32.5	81.7
	C	23	18.3	18.3	100.0
	Total	126	100.0	100.0	

APPENDIX D
THE MEAN STRUCTURE OF THE STUDY VARIABLES

The Mean Structure for Graded Level of Formal School Attended for every semester.

Time=Semester	Graded school A	Graded school B	Graded school C
1	59.5924	58.3895	58.2222
2	56.5322	58.3522	56.0422
3	58.2711	57.0983	56.0930
4	59.8298	58.3859	56.7583
5	56.5957	56.3373	54.9739
6	61.5463	60.0666	57.1722
7	58.7279	58.1049	56.5443

***The Mean Structure* for Geographical Location of Students**

Obs=SWA	Northern Belt(L_1)	Central Belt(L_2)	Eastern/Volta Belt(L_3)	South costal belt(L_4)
1	55.4293	60.2056	58.3536	59.3684
2	54.3140	57.0565	58.6164	56.6187
3	53.8120	57.4096	57.5242	59.3906
4	53.7780	59.4371	59.0564	60.0010
5	54.6380	55.7635	57.3494	56.4842
6	58.2780	58.8635	60.8779	62.7910
7	55.0420	58.1229	58.6297	59.1087

Descriptive Statistical Data Analyses of the Covariance Matrix

Covariance Matrix, DF = 126							
	SWA1	**SWA2**	**SWA3**	**SWA4**	**SWA5**	**SWA6**	**SWA7**
SWA1	44.2896	10.7708	29.1830	14.5995	-3.2280	-0.3503	15.8741
SWA2	10.7708	70.1979	14.6413	12.9442	11.9701	14.4507	22.4979
SWA3	29.1830	14.6413	77.3809	10.6216	9.3714	18.7123	26.6474
SWA4	14.5995	12.9442	10.6216	57.8285	-0.4060	4.2758	16.6407
SWA5	-3.2280	11.9701	9.3714	-0.40601	95.1048	30.3853	23.8704
SWA6	-0.3503	14.4507	18.7123	4.2758	30.3853	62.9426	21.7373
SWA7	15.8741	22.4979	26.6474	16.6407	23.8704	21.7373	21.2107

Descriptive Statistical Analyses for Pearson correlation coefficients matrix.

Pearson Correlation Coefficients, N = 126			Prob > |r| under H0: Rho=0				
	SWA1	**SWA2**	**SWA3**	**SWA4**	**SWA5**	**SWA6**	**SWA7**
SWA1	1.00000	0.19317 0.0296	0.49850 <.0001	0.28848 0.0010	-0.04974 0.5787	-0.00664 0.9410	0.51792 <.0001
SWA2	0.19317 0.0296	1.00000	0.19866 0.0252	0.20316 0.0220	0.14650 0.1003	0.21740 0.0141	0.58305 <.0001
SWA3	0.49850 <.0001	0.19866 0.0252	1.00000	0.15878 0.0746	0.10924 0.2215	0.26813 0.0023	0.65775 <.0001
SWA4	0.28848 0.0010	0.20316 0.0220	0.15878 0.0746	1.00000	-0.00547 0.9513	0.07087 0.4285	0.47514 <.0001
SWA5	-0.04974 0.5787	0.14650 0.1003	0.10924 0.2215	-0.00547 0.9513	1.00000	0.39273 <.0001	0.53147 <.0001
SWA6	-0.00664 0.9410	0.21740 0.0141	0.26813 0.0023	0.07087 0.4285	0.39273 <.0001	1.00000	0.59492 <.0001
SWA7	0.51792 <.0001	0.58305 <.0001	0.65775 <.0001	0.47514 <.0001	0.53147 <.0001	0.59492 <.0001	1.00000