CORRELATED DATA ANALYSIS ON STUDENTS ACADEMIC PERFORMANCES

By

EKOW EWUSI AMISSAH, B.Sc.

THESIS

Presented to the Department of Mathematics,

Kwame Nkrumah University of Science and Technology

in Partial Fulfillment of the Requirements for the Degree of

MASTER OF PHILOSOPHY

College of Science

November, 2012

Declaration

I hereby declare that this submission is my own work towards the Master of Philosophy (M.Phil.) 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, except where due acknowledgement has been made in the text.

Ekow Ewusi Amissah(PG5072610)	EAmissah	04/12/2012
Student	Signature	Date

Certified by:

Nana Kena Frempong		04/12/2012
Supervisor	Signature	Date
Mr. K. F. Darkwah	KFDarkwah	23/1/2013
Head of Department	Signature	Date

Abstract

This study examines different factors that might have impact on the academic performance of undergraduate students other than the Semester Weighted Average (SWA). Data on the SWA of Class of 2012 Mathematics students was acquired from the Quality Assurance and Planning Unit (QAPU) and the Examination Office of the Department of Mathematics, Kwame Nkrumah University of Science and Technology (KNUST). The main factors considered for this research were entry age, gender, entry aggregate, Ghana Education Service (GES) graded level of Senior High School attended and geographical location. The statistical models considered were Generalized Estimating Equation (GEE) and Random Effect Model. From the findings, the effect of entry aggregate means students with a lower entry aggregate (better aggregate) have an approximate 33% increase in the average SWA scores. Students from the *Northern* Belt on an average with regards to SWA score increased approximately by 1.7% per semester, students from the *Middle* and *Southern* Belt increased by 1.2% and 0.9% per semester respectively. On the average, students with lower entry ages have ~~on the~~ average ~~a~~ higher SWA score. The GEE and Random Effect Model showed similar significant factors ie. age , entry aggregate and geographical location.

Table of Contents

Chapter

Appendix

List of Tables

List of Figures

LIST OF ABBREVIATIONS

WASSCE	West Africa Senior Secondary Certificate Examination
SHS	SeniorHigh School
GES	Ghana Education Service
SWA	Semester Weighted Average
CWA	Cummulated Weighted Average
KNUST	Kwame Nkrumah Univerity of Science and Technology
GEE	Generalized Estimating Equation
GLM	Generalized Linear Model
QAPU	Quality Assurance and Planning Unit
SPSS	Statistical Package for Social Sciences
MRM	Mixed-Effect Regression Model
CPM	Covariance Pattern Model
AR	Autoregression
IRLS	Iteratively Reweighted Least Squares
WLS	Weighted Least Squares
MLE	Maximum Likelihood Estimation
RMLE	Resulting Maximum Likelihood Estimation
GES SHS CAT	Ghana Education Service Senior High School Category

Dedication

To all

those who are yet to climb the ladder of Philosophy

...Make hay even when its still raining; for you may not know if it will ever stop raining...

Acknowledgements

Most importantly, my biggest and highest appreciation, goes to the Elohim God my heavenly Father. Your love, grace and favour through Christ is 'massive' and very much appreciated throughout this programme.

I would like to show my gratitude to my supervisor, Nana Kena Frempong, for his supervision and guidance from the initial to the final stages of this thesis. The knowledge we shared through discussions and brainstormings enabled me develop a much better understanding of the subject. It was timely as well as very helpful.

It is a pleasure to thank Rev.William Obeng-Denteh and Mr. Vincent Dedu, all of the Department of Mathematics, KNUST for providing me with the data for my analysis. I also thank Mr. Kojo Ankar-Brewoo of the Quality Assurance and Planning Unit (QUAPU) for providing me with data.

My thanks also go to the Department of Mathematics, KNUST, for permitting me to carry out my research in the department.

I owe my deepest gratitude to my mother, Miss Nancy Apagya-Bonney for supporting and financing me throughout my tertiary education.

I thank my all, Mr. Emmanuel Kwesi Sam and family for all the kind words, finance and continuous encouragement that always pushed me to work hard.

Chapter 1

INTRODUCTION

1.1 Overview

This chapter gives a background of the study, statement of 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.

1.2 Background Studies

Education is the act or process of imparting or acquiring general knowledge, developing the powers of reasoning and judgment and generally of preparing one's self or others intellectually for mature life.

The Ghanaian society of late has placed high value upon successful academic performance. As an up and coming responsible citizen, occupations which command high positions of status in our society are those which have, among other characteristics, extensive and rigorous educational requirements.

Academic achievement has been variously defined as a level of proficiency attained in academic work as formally acquired knowledge in school subjects, which is often represented by percentage of marks obtained by students in examinations (Kohli (1975)). To reach the goal of excellence in the academic sphere and to optimize academic achievement to a maximum, a review of correlated academic achievement and its implications for educationists and policy makers would be meaningful.

Physicians, lawyers, engineers, researchers and those in other professional occupations must achieve the successful completion of arduous educational careers. Considerable emphasis is placed upon superior academic performance as well.

Admission into certain course in the university world wide is attained through a complicated series of admission procedure. In a case whereby 1000 candidates may apply to a single school of Pharmacy or Civil Engineering , as few as 90 may gain regular admission. A very high academic performance in the West Africa Senior Secondary Certificate Examination (WASSCE) at the Senior High School (SHS) level is almost a major requisite for consideration.

Studies of academic performance may be able to explain to some extent some factors that may cause an increase the likelihood of high academic performance and to possibly make recommendations to facilitate the removal of social, psychological and economic barriers to full participation in the educational system by all and sundry who have the ability and appetite to do so.

Some of the impetus for this study came from initiative to help the Department of Mathematics and its administration to monitor and evaluate the gradual progress of students from year 1 to year 4.

Academic performance has been traditionally measured by *grades* as a means of expressing a students scholastic standing. At the present time, grades are the criteria by which our educational system judges scholastics performance. Examination results are direct indicators of a students capacity to cope with university work.

The initial investigations of some researchers in their bid to find out whether there exist certain variables affecting academic performance indicate a position relationship between students achievement in Mathematics and home background variables such as parents level of education, Mathematics students and their Ghana Education Service (GES) grade, type of the SHS attended, gender, entry aggregates, fee paying status, loan financial assistance, literacy status of students, choices of students' optional courses and attitude of students in Mathematics. In addition, the relative effect of the index of socio-demographic factors such as age as well as self-concept affect students academic performance as well as in Mathematics (Papanastasiou (2002)).

Students performance with respect to the Semester Weighted Average (SWA) scores is a topic of great practical concern to lecturers and parents and of great theoretical concern to researchers. Achievement outcomes have been regarded as a function of two characteristics, *skill* and *will* and these must be considered separately because possessing the *will* alone may not insure success if the *skill* is lacking (McCombs and Marzano (1990)).

The SWA is basically a single score that represents a students performance in all the course taken in a semester and is calculated to represent a students quality of performance numerically. It is calculated by multiplying the marks obtained for each course by the credits of the particular course adding up the products and dividing by the total number of unit of credit for the course. To be able to critically access the up to date academic information of a student, Cummulative Weighted Average (CWA) is calculated; which is used for the award of degree. The CWA therefore depends on SWA.

Academic failure is relative concept. According to Good (1973), failure implies lack of success on the part of students in accomplishment of school work. Naturally, it differs as the school work varies; curricula change as well as the standard of assessment. Many factors contribute to academic failure and under-achievement, the major ones being intelligence, personality including motivation and adjustment, home background and school background. Some investigations carried out by researchers indicated that students who have positive perceptions or attitudes towards Mathematics showed better achievement in both Mathematics and Science (Kiamanesh and Kheirieh (2001)); therefore Mathematics can be seen as a tool.

These concept in various multiple variables are related to KNUST Mathematics students SWA score which the study seeks to investigate.

1.3 Problem Statement

One of the main tools used in assessing the students academic performance is the Semester Weighted Average (SWA). The SWA takes into account the performance of students at the end of the semester examinations. Quality Assurance & Planning Unit (QAPU) accesses KNUST students academic performances using the SWA. The question as to what goes into measuring the quality of academic output goes beyond the SWA scores. This thesis seeks to dive deep into this issue at hand.

To achieve quality education, a proper measure of academic performance should be put in place. The basic purpose of any measurement system is to provide an efficient and effective feedback.

Socioeconomic status and certain academic requirements are some of the most researched factors among educational professionals that contribute towards the academic performance of students. The most prevalent argument is that the socioeconomic status of learners affects the quality of their academic performance. Most of the experts argue that the low socioeconomic status has negative effect on the academic performance of students because the basic needs of students remain unfulfilled, making them unable to perform better academically (Adams (1996)). Low socioeconomic status causes environmental deficiencies which results in low self esteem of students(M.Spelling (2005)).

More specifically, this study aims to identify and analyze other factors that may have an impact on students academic performance.

1.4 Objectives

The study attempts to:

1. fit a *Generalised Estimating Equation (GEE) Model* and *Random Effect Model* to the sampled data.

2. compare the models in *objective 1* above.

3. identify the factor(s) that affect academic performances.

1.5 Scope and Limitations

1. This thesis was limited to the class of 2012 Mathematics students.

2. Time and lack of cooperation on the part of some KNUST officials made it difficult retrieve enough data to carry out a comparative study.

1.6 Methodology

This study seek to use Generalized Estimating Equation (GEE) model, which is an extention of Generalized Linear Model (GLM).

The study further uses the Random Effect Model, also known as the variance component model which is hierarchical linear model to analyze the data and its entails.

Graphical representation of each technique of some technique of some stages has beeb displayed with the aid of a statistical software SAS version 9.1, SPSS version 16, Microsoft excel 2010 and R.

Semester Weighted Average (SWA) of Class of 2012 Mathematics students were useful for this thesis.

1.7 Justification of Problem

This study would help Quality Assurance and Planning Unit (QAPU) to improve reported analytical results and also enlighten them on some other factor worth considering. This would go a long way to help produce information and policy implementations.

The University authorites would also be informed about the contribution of significant factors on students' academic performance thereby tracking students records of grades (SWA). This will help improve both teaching and learning process.

The authorities would however be alerted on these significant variables that plays restriction on students' academic performances; thereby influencing certain University policies such as admission process.

1.8 Structure of The Thesis

The following chapters are organized as follows:

- Chapter 1 introduces the research work
- Chapter 2 discusses the literature reviews and other areas of application of the Longitudinal Data, GEE Models and Random Effect in particular.
- Chapter 3 talks about the methodology employed in our study.
- Chapter 4 presents the analysis from the results.
- Chapter 5 presents the discussion, conclusion and recommendations for further research.

Chapter 2

LITERATURE REVIEW

In this chapter, there is a review of research works and concepts of other authors to uncover the academic achievements of students in institutions and the social environment as a whole.

Modeling of repeated categorical data was described by Landis et al.(1977) and provides a large sample asymptotic method that works nicely for data that have adequate sample size, a small number of response points, a small number of categorical explanatory variables measured at the subject level, and no missing data.

Generalized Estimating Equations (GEE) is an essential tool in managing categorical data and it provides a nonlikelihood based approach to modeling repeated or clustered data that applies to a broader set of data situations that are frequently encountered. It handles missing data, continuous explanatory variables, and time-dependent explanatory variables. While responses can be either continuous or categorical, it is especially useful for data that are binary or discrete.

Stewart et al. (1999) in their longitudinal data analyses on the relationship between stress-related measures and academic performance during the first two years of medical school, medical students (n=121) were surveyed prior to beginning of classes and 8 months later variables predisposing to distress, stress response (depression and state anxiety), and stress management strategies were assessed. Pre-medical academic scores and grades at the end of five assessment periods over the course of the first two years of medical school were also obtained. The results shows that academic performances before and during medical school was negatively related to reported stress levels. On the bivariate correlations, there were numerous significant relationships between stress and academic performance.

The purpose of Guay et al. (2004) study was to test children's academic self-concept, family socioeconomic status,family structure (single parent vs. two parent family) and academic achievement in elementary school as predictors of children's educational attainment level in young adulthood within a ten-year longitudinal design. Participants (254 girls, 211 boys) were three cohorts of students in Grades 3, 4, and 5 from ten elementary schools. Results from structural equation modeling revealed that academic self-concept predicted educational attainment level ten years later over and above prior achievement. Moreover, this pattern of results was invariant across cohorts. In addition, regression analyses based on a restricted sample (n = 243) indicated that the academic self-concept/educational attainment level relation was still significant while controlling for family SES, family structure (single parent vs. two parent family), and academic achievement. Discussion focuses on the theoretical and practical implications of the results.

With regards to Marsh and Yeung (1998) longitudinal causal models of growth in Mathematics and English constructs (school grades, standardized tests, academic self-concept, affect and coursework selection) were based on three waves of data from the large (N = 24,599), nationally representative National Education Longitudinal Study of 1988. Math and English self-concepts had significant path coefficients leading to subsequent school grades, courseworkselection, and standardized testscores. Unlike previous studies that did not consider Mathematics and English constructs in the same model, we found these relations to be very domain specific (e.g., there were significant positive paths from Mathematics self-concept to subsequent Mathematics outcomes but not to subsequent English outcomes). Girls had higher scores for all English constructs and Mathematics school grades, but they had lower Mathematics self-concepts. Whereas similar studies conducted over the past 20 years found diminishing gender differences, these data show relativegains for girls in achievement and coursework selection for both Mathematics and English. Path coefficients relating prior math and English constructs to subsequent outcomes, however, were similar for boys and girls. Hence, the extreme domain specificity of relations between prior self-concept and subsequent outcomes was similar for boys and girls.

Previously published research has not moved beyond studying the general association between retention and high school dropout. In this Jimerson et al. (2002) longitudinal study seeks to evaluate within-group differences, exploring the characteristics of those students who are retained and subsequently drop out as compared to those students who are retained and do not drop out. A transactional-ecological view of development is presented to assist in situating the findings within a framework of long-term outcomes across development. The results of this study suggest that there are early socio-emotional and behavioral characteristics that distinguish which retained students are most likely to drop out of high school. In addition, maternal level of education and academic achievement in the secondary grades were also associated with high school graduation status. These findings provide information that extend beyond the association between grade retention and later drop-out. In particular, this investigation suggests that it is especially important to attend to the socio-emotional and behavioral adjustment of children throughout their schooling to facilitate both their immediate and long-term academic success. It has been established that there is a strong connection between high school dropout and grade retention Jimerson et al. (2002). This current longitudinal study moves beyond generalities to examine specific behavioral and academic variables of retained students in order to increase our understanding of what places children at risk for later high school dropout. Both retained students and dropouts present a variety of profiles; however, certain early characteristics may increase the possibility that a retained student will drop out. This longitudinal study is the first to explore characteristics associated with those students who are retained and drop out, in contrast to those who are retained and continue on to graduate from high school. While many studies have demonstrated the strong association between grade retention and dropout, no studies to date have examined within-group characteristics of retained students to explore processes that may provide further understanding of this association. This 12-year longitudinal study provides information addressing the following questions:

1. Do family characteristics differentiate which retained students are more likely to drop

out Maternal level of education and value of education will be compared between those retained students who drop out and those who persist during 11th grade.

2. Do socio-emotional and behavioral characteristics differentiate which retained students are more likely to drop out Socio-emotional and behavioral adjustment in kindergarten, 2nd grade, and 8th grade will be compared between those retained students who drop out and those who persist during 11th grade.

3. Do achievement characteristics differentiate which retained students are more likely to drop out Academic achievement in 2nd, 4th, 5th, 7th, 8th, 9th, 10th, and 11th grades will be compared between retained students who drop out and those who persist during 11th grade.

According to Caprara et al. (2011) personal determinants of academic achievement and success have captured the attention of many scholars for the last decades. Among other factors, personality traits and self-efficacy beliefs have proved to be important predictors of academic achievement. Openness and academic self-efficacy at the age of 13 contributed to junior high-school grades, after controlling for socio-economic status (SES). Junior high-school grades contribute ~~to~~ academic self-efficacy beliefs at the age of 16, which in turn contributed to high-school grades, over and above the effects of SES and prior academic achievement. In accordance with the posited hypothesis, academic self-efficacy beliefs partially mediated the contribution of traits to later academic achievement. In particular, conscientiousness at the age of 13 affected high-school grades indirectly, through its effect on academic self-efficacy beliefs at the age of 16. These findings have broad implications for interventions aimed to enhance children's academic pursuits. Whereas personality traits represent stable individual characteristics that mostly derive from individual genetic endowment, social cognitive theory provides guidelines for enhancing students efficacy to regulate their learning activities.

A sample of 370 students in the 7th and 9th grades in 1998 in Scalesa et al. (2006) was followed for 3 years through the 10th and 12th grades in order to investigate the relation of developmental assets positive relationships, opportunities, skills, values, and self-perceptions to academic achievement over time, using actual GPA as the key outcome variable. The greater the number of developmental assets students reported in the 7th - 9th grades, the higher their GPA in the 10th-12th grades. Students who stayed stable or increased in their asset levels had significantly higher GPAs in 2001 than students whose asset levels decreased. Increases in assets were significantly associated with increases in GPA. Experiencing in 1998 clusters of specific assets increased by 2-3 times the odds of students having a B+ or higher GPA in 2001. The results offer promising evidence that a broad focus on building the developmental nutrients in young people's lives may contribute to academic success. The results of the present study offer promising evidence of how a broad focus on building the developmental nutrients in young people's lives may contribute to promoting their academic achievement. Building developmental assets thus merits consideration as one of the strategies districts and communities can use to positively affect achievement. Especially if careful attention is paid to how strategies for promoting this positive youth development can be infused to strengthen classroom practices and curriculum and instruction then the already apparent link between developmental assets and achievement likely can be strengthened.

The associations between maternal labour force participation and child academic achievement were examined according to Horwood and Fergusson (1999) in a birth cohort of New Zealand children who have been studied from birth to age 18. The results of this analysis suggested the presence of small associations between the extent of maternal labour force participation and scores on standardised tests of word recognition, reading comprehension, and mathematical reasoning. Similar associations were found between maternal labour force participation and success in school leaving examinations. These associations arose predominantly because children whose mothers worked had better performance than children whose mothers who had not worked in paid employment. However, patterns of maternal labour force participation were also related to a series of family and child factors including : maternal education, family socioeconomic status, race, birth order, family composition, early mother child interaction, and child IQ. Adjustment for these factors reduced associations between maternal labour force participation and academic achievement to the point of practical and statistical nonsignificance. These results were found to be robust and similar conclusions were found for :

- a range of measures of maternal labour force participation
- subgroups of the cohort defined by gender, socioeconomic status, ethnicity, or family

type

It has been found in McLemore (2010) that the juncture of three domains allows effective learning to take place. First, students need to be able to receive and process information, a process known as cognition1. Second, students need to be able to physically communicate their knowledge, a process relating to their psychomotor skills.And third, students need to possess affect, the desire to learn and understand why their participation is beneficial to them (this is also called social emotional learning).A growing body of research points to resilience, a key component of social emotional learning, as being a critical facet of education. Referring to the ability to succeed in school despite adverse conditions such as poverty or abuse, resilience includes components such as confidence, a sense of well-being, motivation, an ability to set goals, relationships/connections, and stress management. Resilience, combined with the creation of positive and constructive learning environments, can benefit all students, regardless of their risk level, and can effectively be taught in schools. The development of these skills is key in improving academic performance and enabling students to experience success in school, and subsequently, in life.

To what extent and which personality traits predict academic performance was investigated according to Chamorro-Premuzic and Furnham (2003) in two longitudinal studies of two British university samples. Academic performance was assessed throughout a three years period and via multiple criteria (e.g., exams and final-year project). In addition several indicators of academic behaviour, e.g., absenteeism, essay writing, tutors exam predictions, were also examined with regard to both academic performance and personality traits. In sample 1 (N 1/4 70), the Big Five personality factors Costa and McC-rae (1992) particularly Neuroticism and Conscientiousness were found to predict overall final exam marks over and above several academic predictors, accounting for more than 10 percent of unique variance in overall exam marks. Results suggest that Neuroticism may impair academic performance, while Conscientiousness may lead to higher academic achievement. In sample 2 (N 1/4 75) the EPQ-R, Eysenck and Eysenck (1985) was used as the personality measure and results showed the three superfactors were the most powerful predictor of academic performance, accounting for nearly 17 percent of unique variance in overall exam results. It is demonstrated that (like Neuroctisim) Psychoticism could limit academic success. The present results provide evidence supporting the inclusion of well established personality measures in academic selection procedures, and run counter to the traditional view of ability measures as the exclusive psychometric correlate of academic performance.

A cohort of students took five chemical engineering courses taught by the same instructor in five consecutive semesters according to Felder et al. (1995). This report examines gender differences in the students academic performance, persistence in chemical engineering, and attitudes toward their education and themselves. The women in the study on average entered chemical engineering with credentials equal to or better than those of the men, but exhibited erosion relative to the men in both academic performance and confidence as they progressed through the curriculum. The backgrounds and pre-engineering academic credentials of the women in this study marked them as more likely to succeed than the men. Their parents were on average more highly educated and had received more training in science or technology, they scored equally well or better than the men on pre-college admission tests, and Learning and Study Strategies Inventory results indicated that they were more highly motivated to study and made better use of study aids. This report examines gender differences in the student's academic performance, persistence in chemical engineering, and attitudes toward their education and themselves. The women in the study on average entered chemical engineering with credentials equal to or better than those of the men, but exhibited erosion relative to the men in both academic performance and confidence as they progressed through the curriculum. The men in the study consistently earned equal or higher grades in chemical engineering courses than did the women, and the percentage of men earning A's in several courses was significantly greater than the percentage of women doing so. Of the students who had intended to major in chemical engineering when they began the first course, the percentage of women who dropped out for any reason after the sophomore year was twice the percentage of men dropping out. The percentages dropping out by the end of the senior year were closer, with relatively more women transferring into other curricula and considerably more men dropping out of school or being suspended. Throughout the period of the study, men who failed a chemical engineering course were more likely than women to repeat the course and remain in the curriculum, while women who failed a course were more likely than the men to switch out of chemical engineering. These results are consistent with patterns described by Dweck and Repucci; Dweck C. and Repucci N. (1973) who observed that young men are more likely than young women to persist in the face of academic challenges, and also noted a greater tendency of women than men to drop out of engineering.

From Umar et al. (2010) social factors affect academic performance in terms of time demand and the psychological state, these came as a result of their study on social effort on academic performance of student considering some social factors such as romantic relationships, organizations and clubs and sports activities. The immidiate question that comes out is how one strikes a balance between the stressful academic attainment and social activities. According to them all these factors have a direct or indirect relationship with students' performance. Their research found out that student cults are an academic impediment, romantic relationships have the highest impact and may be a psychological barrier to an effective learning process also excessive sporting activities and involvement in clubs and organizations may pose a threat, but an insignificant one.

Paulo S.(2006) also researched about growth status on academic achievement using multiple regression model analysis and univariate analysis after assessing academic performance and measuring growth using standard procedure and height-for-age of 277 student selected randomly from his department. He came out that student whose growth move with their height and age perfrom better than those who do not grow well with their height and age (have retarded growth), which indicates that retardation have a negative impact on academic performance.

Erdinc C. (2005) investigated the effort of gender and efficiency on academic performance on university grade level stated that self-efficiency has been define as an individual's judgement of his or her capability to organize and execute the course of action required to attain designated types of performances. To him strong sense of efficiency increase human accomplishment and people with high connfidence in their capabilities approach difficult tack as challenge to mastered rather than as threats to be avoided. Using stochastic rasch modeling came out with this conclusion, there was significant relationship between personal efficiency and academic performance and there was no significant effect on gender on academic performance.

Rainfall can be modelled using various timescales including hourly, daily, monthly and annual timescales Dunn P. (2004). The most commonly studied timescale is daily because it can discriminate between the number of wet days and rainfall amounts when it does rain. This provides a more detailed understanding of the rainfall process Chandler and Wheater (2002). Little research has been completed on annual or monthly rainfall since 1985, Srikanthan and McMahon (1985). Annual rainfall models have little direct application, however they are used in disaggregation schemes to obtain monthly data, which is used in the estimation of water demand and the simulation of water supply systems. Generalized Linear Models (GLMs) are also becoming increasingly popular to model rainfall data. GLMs are also the backbone of Generalized Estimating Equations (GEE), and GEE's are the main focus of this workpiece.

The GEE seeks to design and analyse the population-average models in Preisser et al. (2003) for randomized and non-randomized cluster trials in an integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials . A design strategy that selects matching control communities based upon propensity scores, a statistical analysis plan for dichotomous outcomes based upon Generalized Estimating Equations (GEE) with a design-based working correlation matrix, and new sample size formulae are applied to a large non-randomized study to reduce underage drinking. The statistical power calculations, based upon Wald tests for summary statistics, are special cases of a general power method for GEE.

A cluster-randomized control trial was used in Loyalka et al. (2010) to examine the effects of providing information on college costs and financial aid to high school students in poor regions all over the world. Some students receive some types of financial aid, but has no significant effect on specific college choices or whether students persist to go to college. Analyses in cluster-RCTs with binary outcomes also may use generalized estimating equations (GEE) models. The GEE model adjusts the correlation matrix used in the estimation of the treatment effect instead of including a cluster-level error term Liang and Zeger (1986). More importantly, the GEE model estimates the average effect of the intervention across the population rather than a cluster-specific intervention effect.

Profit-sharing as a means of increasing productivity, employment, individual earnings and labor market flexibility has been discussed in Pannenberg and Spiess (2007) among politicians as well as economists for a long time. Generalized Estimating Equations (GEE) approach is also used to analyse the mean and the covariance structure of a bivariate time series process of panel data with mixed continuous and discrete dependent variables. The approach is used to jointly analyze wage dynamics and the incidence of profit-sharing in West Germany. Through this approach, permanent unobserved individual's ability is comparatively more important in the profit-sharing than in the wage equation. It also shows that shocks have a long-lasting effect on transitory wages but not on the incidence of profit-sharing. Hence, the results support theoretical predictions that selection into profit-sharing is mostly due to unobservable ability and that profit-sharing ties wages more closely to productivity.

As part of the research into Structural Conservation Practices in U.S. Corn Production in Schaible et al. (2007), working-lands programs assist farmers in implementing and maintaining such land-management and structural practices such as conservation tillage, crop rotatio, cover crops, enhanced nutrient management, precision agriculture, irrigation water management, crop/livestock diversity, and the use of infield and perimeter-field structures such as strip cropping terraces, and stream-side herbaceous buffers. In an effort to better understand the relationships between farmer motivations, program incentives, and the environmental benefits of conservation programs, USDA also initiated the pilot national survey integration program, the Conservation Effects Assessment Project - Agricultural Resources Management Survey (CEAP-ARMS). CEAP-ARMS integrates National Resource Inventory (NRI) data on field-level physical characteristics and CEAP production practice and program participation information with USDA ARMS data on cost-of-production, operator, farm household, and farm resource/economic data. CEAP-ARMS was completed in 2004 for wheat. A Generalized Estimating Equations (GEE) procedure is used to estimate two models. The cost-function models estimate field-level, produceacreage allocation decisions, first, as a function of normalized production input prices (Model I), and second, as a function of normalized input prices and several key exogenous variables reflectinthe potential influence of a variety of field, farm, and environmental characteristics (Model II).Because the dependent variable in this analysis is continuous, we use a Generalized Estimating Equations (GEE) procedure to estimate two models. The GEE estimation procedure Liang and Zeger (1986) accounts for the correlation between adoption decisions measured as a continuous variable, while maintaining the theoretical integrity of a multinomial discrete-choice model typically used in technology adoption studies. The two cost-function models estimate field-level, producer acreage allocation decisions for corn, first, as a function of normalized production input costs (prices) and structural technology class and installation time-period attributes (Model 1), and second, as a function of Model 1 variables plus socio-environmental variables reflecting the potential influence of a variety of field, farm, and environmental characteristics (Model 2). Estimated GEE coefficients for Model I and their significance tests indicate that relative prices do explain producer choices in allocating field acres between corn production, and infield and perimeter-field conservation structural practices .

Tek K. et al. (2009) illustrated in the analysis of longituidinal data using GEE and also shows how output from SAS macros can be streamlined and organized to aid interpretation of analysis. They made it clear that the use of GEE through procedures such as SAS PROC GENMOD is becoming increasingly common place, as far as model evaluation is concerned. It's widespread use is somehow limited by the lack of easily accessible measures to evaluate the model goodness -of-fit directly from the default SAS output. The study came up with three goodness-of-fit indices in a SAS macro, using various working correlation matrices for model comparison. Their workpiece illustrates with a longitudinal data set, how four models with different working correlation matrices specification with a binomial logit link function are generated using macros. The results suggests that the estimated coefficients for the models were largely similar. They therefore agree with Zeger and Liang (1986) point that misspecification of working correlation would still give consistency result. Their study also illustrates the procedure of data management and preliminary data analyses work needed before carrying out similar analyses using several simple SAS macros.

The random effect model is alo used in the analysis of hierarchical or panel data when one assumes no fixed effects. Random effects by Gutierrez R. (2008) are not directly estimtated but summerized by their variance components, which are estimated from the data.

Muhammed A. et al. (2010) as part of their reseach workpiece used only six year data to find the causal relation between price to earnings ratio (P/E) taken as a measure of corporate perfromance and cost of equity used random effect model in analysing the causal relation between dependent and independent variables. The study made it clear that there is a weak relationship between price to earnings and cost equity.

Mandel M. and Betensky R. (2008) study in estimating time-to-event from longitudinal ordinal data using radom-effect Markov models made it clear that Longitudinal ordinal data are common in many scienctific studies including those of multiple sclerosis (MS), and are frequently modeled using Markov dependency. They said several authors have proposed radom-effects Markov models to account for heterogeneity in the population. In their paper, they went one step further and study prediction based on random effects Markov models in particular. They managed to show how to calculate the probabilities of future events and confidence intervals for those probabilities, given observed data on the ordinal outcome and a set of covariates, and how to update them over time.

Stinebrickkner R. and Stinebrickner T. (2007) concluded that very small amount of existing work has provided direct evidence about the reationship between studying and academic performance. As part of their research they collected measures of studying-effort and obtained estimates of the (conditional) correlation between the number of hours that a person studies and their corresponding academic performance.

In Rodriquez C. (2006), the reseacher seek to analyze the relations among the levels of skills in the technology handling (ICT Skills Index), educative and professional level of the father of the student, surrounding or the environment in which the student live, in order to propose a model of causal relations that represents suitably, the effects of these factors use with academic aims ~~on~~ the results in the studies. It was a random, cross-sectional and anonymous study: they begin by being a descriptive research and finishes as explanatory study. The statistical analysis used was carried out in two phases: the descriptive univariado and with structural equations modelling. The results shows that the education and professional level of the father have the strongest effect on academics. Surrounding or the environment in which the student live and individual behaviour was the second, ICT skills index had significant effect on academic performance.

Henry M. and Supovid D. (2004) stated that education is a cummulative process. According to them while students' knowledge and skills are built up over time, educational researches are not often afforded the opportunity to examine the effects of interventions over multiple years. In their study about America's Choice school reform design which is just an opprotunity. Using a 11 year longitudinal data of student's academic performance from Rochester, NY they examine the effects of America's Choice on students learning gains from 1998 to 2003. to analyze these longitudinal data, they used an advanced sophisticated statistical modeling technique called Bayesian hierarchical growth curve analysis with crossed random effects. This technique enabled them to model the annual growth in individual's academic performance.

Longitudinal panel data examples in Rogosa and Saner (1995) are used to illustrate estimation methods for individual growth curve models. These examples constitute one of the basic multilevel analysis settings, and they are used to illustrate issues and concerns in the application of hierarchical modeling estimation methods,specifically, the widely advertised HLMprocedures of Bryk ~~and~~ Raudenbush. One main expository purpose is to demystify these analyses by showing equivalences with simpler approaches. Perhaps more importantly, these equivalences indicate useful data analytic checks and diagnostics to supplement the multilevel estimation procedures. In addition, we recommend the general use of standardized canonical examplesfor the checking and exposition of the various multilevel procedures; as part of this effort, methods for the construction of longitudinal data examples with known structure are described.

A model in Diggle and Kenward (1994) is proposed for continuous longitudinal data with non-ignorableor informative drop-out (ID). The model combines a multivariate linear-model forthe underlying response with a logistic regression model forthe drop-outprocess. The latter in corporates dependence of the probability of drop-out unobserved,or missing,observations.Parameters in the model are estimated by using on maximum likelihood (ML) and inferences drawn through conventional likelihood procedures. In of likelihood ratio tests can be used to assess the informativeness the drop-outprocess particular,to through comparison of the fullmodel with reduced models corresponding randomdrop-out(RD) and completely random processes. A simulation study is used to assess the procedure in two settings:the comparison of time trends under a linearregressionmodel with autocorrelated errors and the estimation of period means and treatment differences from a four-period four-treatment crossover trial. It is seen in both settings that,when data are generated underan ID process, the ML estimators from the ID model do not suffer from the bias that is present in the ordinaryleast squares and RD ML estimators.The approach is then applied to three examples. These derive from a milk protein trial in volving three groups of cows, milk yield data from study of mastitisin dairycattle and data from a multicentre clinical trial on the study of depression. All thre eexamples provide evidence of an underlying process, two with some strength. It is seen that the assumption of an ID rather than an RD process has practical implications for the interpretation the data.

This paper Zajacova et al. (2005) investigates the joint effects of academic self-efficacy and stress on the academic performance of 107 nontraditional, largely immigrant and minority, college freshmen at a large urban commuter institution. We developed a survey instrument to measure the level of academic self-efficacy and perceived stress associated with 27 college-related tasks. Both scales have high reliability, and they are moderately negatively correlated. We estimated structural equation models to assess the relative importance of stress and self-efficacy in predicting three academic performance outcomes: first-year college GPA, the number of accumulated credits, and college retention after the first year. The results suggest that academic self-efficacy is a more robust and consistent predictor than stress of academic success.

This paper Gibson and Cassar (2005) investigates causal relationships between planning and performance utilizing a longitudinal database with responses from 2,956 businesses over the four-year the association between period. Results confirm that is evident in most planning activity and performance extant literature.They also, however, cast doubt on the of traditional perception the causal sequence of that association. Although subject to a number limitations, results indicate that planning is more likely to be introduced into an after period of growth rather than before a period small firm of growth.These results make an important contribution to the for understanding planning performance relationship two main reasons: they overcome the static data and relatively of smaller sample size restrictions many past studies; and,they the provide evidence concerning sequence of the relationship between planning and performance.

This article Schneider and Lee (1990) reviews the findings of a field-basedstudy that comparedthe academic performance East Asians and Anglo elementary schoolstudents. Variations in academic performance viewed as the result of the relationship are between socio-cultural factors and interpersonal interactions. Results link the academic success of East Asian students to the values and aspirations they share with their parents, to the home learning activities, in which they participate with their families, and to the expectationsand interactions they have with their teachersand classmates. Results of this study indicate that East Asian academic success is related to cultural and socioeconomic characteristics, and interactive relationships among children, parents, teachers, and peer groups. The use of a holistic research methodology made it possible to uncover several important explanations for why East Asians succeed in school.From an academic perspective, the findings are very positive: Asians do well in school because their parents expect it, their teachers expect it, and their peer group expects it. Parent expectations are extremely powerful and are transmitted through a cultural context in which education is highly valued because it leads to self-improvement and increases self-esteem. But it is not expectations alone that contribute to Asian academic success. Parents help Asian students succeed by carefully structuring their out-of-school time so it is directed at academic related skills. Furthermore, the quiet, industrious, disciplined, and orderly behaviors emphasized in East Asian cultures are rewarded at school, where teachers tend to interpret these behaviors as traits of a good cooperative student.

Researchers have been challenged to go beyond socioeconomic status in the search for school-level characteristics that make a difference in student achievement. The purpose of the present study in Hoy et al. (2006) was to identify a new construct,academic optimism, and then use it to explain student achievement while controlling for socioeconomic status, previous achievement,and urbanicity. The study focused on a diverse sample of 96 high schools. A random sample of teachers from each school provided data on the school's academic optimism, and student achievement scoresand demographic characteristics were obtained from the state department of education. A confirmatory factor analysis and hypothesis tests were conducted simultaneously via structural equation modeling. As-predicted, academic optimism made a significant contribution to student achievement after controlling for demographic variables and previous achievement. The findings support the critical nature of academic optimism.

Teachman (1997) used data on twins to obtain estimates of the degree of shared familial influence on cognitive achievement and academic performance. The results indicate that there is considerable shared influence on cognitive achievement but much less for academic performance.The model estimated also indicates a considerable degree of sibling symmetry in both outcomes. The only significant difference rests between opposite gender twins. For these twins, nonshared sources of influence are more important than for same gender twins. Finally, the results indicate that the regression of academic performance on cognitive achievement is not biased by unmeasured familial characteristics.With that, the results indicate considerable symmetry in both cognitive achievement and academic performance for sets of twins distinguished according to gender composition. For cognitive achievement, there are no variations in the underlying factor structure. Factor loadings and error variances are uniform across types of twin pairs. Similarly,there are no differences in the way twins within a family respond to the common family environment for cognitive achievement. Even though only a single indicator of academic performance (grades) was available, there are no differences in error variances within twin pairs and no differences in the way siblings within a family respond to the common family factor for academic performance.The results also indicate that there is substantially more between-family variation in cognitive achievement than academic performance. That is,twin pairs are more alike with respect to the former than the latter. This finding implies that shared family factors are more important in generating differences in cognitive achievement than academic performance.

Results from surveys generally conclude that they are are, much like other professionals. But high satisfaction registered in surveys may measure the normatively-generated, public side of work that affirms the academic profession'sofficial image. Based on a recent national study of academic physicists' careers, this paper presents results from in-depth interviews in which respondents at a range of US universities provided detailed accounts of their experience in, and identification with academics. Jhermanowicz (2003) study satisfaction from a different angle - through the self-doubts scientists have about their work and careers- and investigate how selfdoubts may systematically differ across distinct 'social worlds' of the academy. First, self-doubts inform our understanding of the social worlds of academe. The academic worlds set self-doubts apart in systematically different ways. Self-doubts connected to work are most apparent among elites, who doubt progress, standing and status. Such self-doubts are apparent but less pervasive among pluralists and are apparent only in anomalous communitarians. In objectiveterms, therefore, further 'up' one that goes in the institutionalhierarchy moving towardthe elite end of the continuum-the more one finds individuals with the same anxieties:doubts tied to work get more common and more intense among higher status people. The elite world appears to induce self-doubts in individuals-because of the competitiveness over recognition that permeates this environment. Thus, conceptualizing the profession in terms of social worlds that are differentiated psychosocially enables us more effectively to understand the logic of satisfaction obtaining in any one of its ecological regions.Second, the use of interviews (and particularly the emphasis on retrospectively- and prospectively-grounded career narratives) has allowed us to look at satisfaction in a more developmental frame. That is, the type of data have allowed us to see how people themselves arrive at a sense of self-understanding that is not fixed in time, but rather accounts for the courses they have travelled, in the work contexts where those travels have been socially situated. This is a significant advantage, since satisfaction, as it is known to people who experience it, is a fluid aspect of work that is responsive to trajectories, turning points and contingencies of a career.The findings suggest that satisfaction is a more nuanced component of work than previous studies have suggested. Explore satisfaction as a developmental process in which people learn how to narrate their careers in ~~the~~ socially accepted formats that their given world of academic work prescribes.

Shachar M. and Neumann Y. (2003) in their research on student's performance consider distance students (DS) and regular students (RS). Using meta-analysis they use their final year grade/score within the two year period to analyze their performance. To them 86 students (N=86) met the established criteria for meta-analysis. In their view, a meta-analysis on a given research topic is directed toward the quantitative integration of findings from various studies. The data extraction and analysis from these works produced 86 calculated effect sizes which yielded the final academic factor. These 86 effect sizes were the basis for the meta-analysis iterations conducted; DS have their mean at 50th percentile while RS have their mean at 65th percentile. They came with the conclusion that their null hypothesis defined as there is no difference between DS and RS instruction for the final academic performance factor should be rejected in favour of our alternative hypothesis.

2.1 FACTORS AFFECTING STUDENTS ACADEMIC PERFORMANCE.

This workpiece seek to manage correlated data on student's academic performance. For some time now, the Ghanaian society have placed high value upon successful academic performance. Occupations which command high positions of status in our society are those which have, among other characteristics, extensive and rigorous educational requirements. Academic performance refers to how students deal with their studies and how they cope with or accomplish different tasks given to them by their tutors. Success in every 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.

Physicians, lawyers, reseachers and those in other professional occupations must achieve the successful completion of arduous educational careers. Many researchers proceed to draw conclusions about the individuals who have opted out of the educational system or who have, for various reasons been excluded. Therefore, studies of academic performance may be able to explain to some extent what factors increase the likelihood of high academic performance, and to possibly make recommendations to facilitate the removal of social, psychological and economic barriers to full participation in the educational system by all those who have the ability and desire to do so.

Parents care about their child's academic performance because they believe that 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 conserns about the school's reputation and the possibility of onetary aid from government institutions, which can hinge on the overall academic performance of the school. Departments of education are charged with improving schools, and so devie method of measuring success in order to create plans for improvement.

Formally, 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 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 the SWA or CGPA is an academic assessment tool used for striking the academic average performance of students achieved in his or her placed institution (Bell, 2010)

Performance in schools 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 performances on standardized tests geared towards spicific ages and based on a set of achievements students in each age group are expected to meet.Certain factors, perhaps beyond control, can impinge restriction of students' academic performance in getting a high SWA. A couple of these factors are:

2.1.1 Age Determinant Variable

The age of student is also correlated with maturity and motivation, which has been shown to be a good predictor of academic performance (Evans, 1999). Student who take some time off before starting tertiary study will generally be older in their first year of study than a student who pregresses directly to tertiary study after leaving school. Students who take a gap-year out-perform student who progress directly. Younger students are more likely to complete qualifications than older students in actual terms, but older students generally out-perform their younger counterparts when controlling for full-time and part-time study status (Scott and Smart, 2005).

2.1.2 Gender Issue

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

All of the research reviews support the hypothesis that student performance depends on different factors including gender. 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 students' performance (Evans (1999)).

Winston et al. (2002) focussed on student's impatience (his time-discount behaviour) 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 especially 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) results. Sacerdote (2001) finds that grades are higher when students have unusually academically strong room mates.

2.1.3 Geographical Location and background of students

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)).

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

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 and Miller (2004)).

2.1.4 Grade level of SHS attended

In Ghana, GES has categorized SHS according to availability of facilities, geographical location, subjects offered and vacancies available to allow candidates to 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'.Apart from partitioning them into stages and levels, it can also be put into day and boarding schedules based on facilities available or the huge number intake. Students who enrolled in the boarding schools often focus in learning relative to academic performance especially when they reach tertiary level.

Chapter 3

METHODOLOGY

3.1 LONGITUDINAL DATA

Change is inevitable. Change is constant.

Benjamin Disraeli

Change is pervasive in everyday life. Infants crawl and walk, children learn to read and write, the elderly become frail and forgetful. Beyond these natural changes, targeted interventions can also cause change: cholesterol levels may decline with new medication; test scores might rise after coaching. By measuring and charting changes like these, both naturalistic and experimentally induced, we uncover the temporal nature of development.

The investigation of change ~~has~~ fascinated empirical researchers for generations. Yet it is only since the 1980s, when methodologists developed a class of appropriate statistical models known variously as individual growth models, random coefficient models, multilevel models, mixed models, and hierarchical linear models that researchers have been able to study change well.

The 1960s and 1970s were especially rancorous, with most methodologists offering little hope, insisting that researchers should not even attempt to measure change because it could not be done well (Bereiter, 1963); (Linn and Slinde, 1977). For instance, in their paper, How should we measure change? Or should we?, (Cronbach and Furby, 1970) tried to end the debate forever, advising researchers interested in the study of change to frame their questions in other ways. Today we know that it is possible to measure change, and to do it well, if you have longitudinal data (Rogosa, Brandt, and Zimowski, 1982); (Willett, 1989).

Longitudinal data or cross-sectional time series data, also called panel data, are data where same entities (panels) like people, firms, and countries were observed at multiple time points. National Longitudinal Survey is an example of panel data, where a sample of people were followed up over the years. On the other hand, General Social Survey data, for example, are not longitudinal data because although a group of people were surveyd for multiple years, the respondents are not necessarily the same each year.

In an attempt to provide a more general treatment of longitudinal data, with more realistic assumptions regarding the longitudinal response process and associated missing data mechanisms, statistical researchers have developed a wide variety of more rigorous approaches to the analysis of longitudinal data. Among these the most widely used include mixed-effects regression models (Laird and Ware, 1981),and Generalized Estimating Equation (GEE) models (Liang and Zeger (1986);Zeger and Liang (1986)). Variations of these models have been developed for both discrete and continuous outcomes and for a variety of missing data mechanisms. The application of these models to the analysis of clustered and or longitudinal data is the primary focus of this workpiece. The primary distinction between the two general approaches is that mixed-effects models are full-likelihood methods and GEE models are partial-likelihood methods. The advantage of statistical models based on partial-likelihoodis that (a) they are computationally easier than full-likelihood methods, and (b) they generalize quite easily to a wide variety of outcome measures with quite different distributional forms. The price of this flexibility, however, is that partial likelihood methods are more restrictive in their assumptions regarding missing data than their full-likelihood counterparts. In addition, full-likelihood methods provide estimates of person-specific effects (e.g., person-specific trend lines) that are quite useful in understanding inter-individual variability in the longitudinal response process and in predicting future responses for a given subject or set of subjects from a particular subgroup (e.g., a county, or a hospital, or a community). Therefore, in managing this kind of data (correlated longitudinal and clustered), those models generally classified as Generalized Estimating Equations or GEE is one of the options to consider as a statistical models for analysis of longitudinal data in this research.

3.2 LONGITUDINAL METHODS

With the development of (new) statistical techniques, such as GEE and random effects analysis, it has become possible to analyse longitudinal relationships using all available longitudinal data, without summarizing the longitudinal development of each subject into one value. The longitudinal relationship between a continuous outcome variable Y and one or more predictor variable(s) X can be described by Equation (3.1).

$$Y_{it} = \beta_0 + \sum_{j=1}^{J} \beta_{1j} X_{itj} + \varepsilon_{it} \tag{3.1}$$

where Y_{it} are observations for subject i at time t, β_0 is the intercept, X_{ijt} is the independent variable j for subject i at time t, β_{1j} is the regression coefficient for independent variable j , J is the number of independent variables, and ϵ_{it} is the error for subject i at time t. This model is almost the same as a cross-sectional linear regression model, except for the subscripts t. These subscripts indicate that the outcome variable Y is repeatedly measured on the same subject (i.e. the definition of a longitudinal study), and that the predictor variable X can be repeatedly measured on the same subject. In this model the coefficients of interest are β_{1j} , because these regression coefficients show the magnitude of the relationship between the longitudinal development of the outcome variable (Y_{it}) and the development of the predictor variables (X_{ijt}). The first extension to this model is the addition of a time indicator t (Equation (3.2)).

$$Y_{it} = \beta_0 + \sum_{j=1}^{J} \beta_{1j} X_{itj} + \beta_2 t + \varepsilon_{it} \tag{3.2}$$

where Y_{it} are observations for subject i at time t, β_0 is the intercept, X_{ijt} is the independent variable j for subject i at time t, β_{1j} is the regression coefficient for independent variable j , J is the number of independent variables, t is time, β_2 is the regression coefficient for time, and ε_{it} is the error for subject i at time t. This simple model can be extended to a general form, in which a correction for both time-dependent covariates (Z_{ikt}) and time-independent covariates (G_{im}) is modelled (Equation (3.3)).

$$Y_{it} = \beta_0 + \sum_{j=1}^{J} \beta_{1j} X_{itj} + \beta_2 t + \sum_{k=1}^{K} \beta_{3k} Z_{ikt} + \sum_{m=1}^{M} \beta_{4m} G_{im} + \varepsilon_{it} \tag{3.3}$$

where Y_{it} are observations for subject i at time t, β_0 is the intercept, X_{ijt} is the independent variable j for subject i at time t, β_{1j} is the regression coefficient for independent variable j , J is the number of independent variables, t is time, β_2 is the regression coefficient for time, Z_{ikt} is the time dependent covariate k for subject i at time t, β_{3k} is the regression coefficient for time dependent covariate k, K is the number of time-dependent covariates, G_im is the time-independent covariate m for subject i , β_{4m} is the regression coefficient for time-independent covariate m, M is the number of time independent covariates, and ε_{it} is the error for subject i at time t. In this general model, again the coefficients of interest are $\beta_1 j$, because these regression coefficients express the relationships between the longitudinal development of the outcome variable Y_{it} and the development of different predictor variables X_{ijt}. Predictor variables and covariates can be either continuous, dichotomous or categorical. For the latter, the same procedure as in cross-sectional linear regression analyses has to be followed, i.e. dummy variables must be created for each of the categories. In the following sections two sophisticated methods (GEE and random coefficient analysis) will be discussed. Both techniques are highly suitable for estimation of the regression coefficients of the general model given in Equation (3.3).

3.2.1 Generalized Linear Model(GLM)

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 represents class of models that are used to fit fixed effects regression models to normal and nonnormal data. McCullagh and Nelders (1989) describe this class of models in great detail and point out that the term "generalized linear model" is due to Nelders and Wedderburn (1972), who indicated how linearity cold 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 variable, and the Poisson regression for count.There are three specification of the GLM. First, the linear predictor, denoted by η, of a GLM is of the form

$$\eta_i = x_i'\beta, \tag{3.4}$$

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 covariate x_i and regression coefficients β. Note that the covariates in x_i can include continuous regressors, 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.

$$g(\mu_i) = \eta_i \tag{3.5}$$

For example, in an ordinary multiple regression, the link function is called the identity link since $g(\mu_i) = \mu_i$ and so $\mu_i = \eta_i$, or

$$E[y_i] = x_i'\beta. \tag{3.6}$$

Under the identity link, the expected value of the dependent variable is simply a linear function of the explanatory varaiables multiplied by their regression coefficients. For dichotomous outcomes, logistic regression Hosmer and Lemeshow (2000) is a popular choice for analysis. this model is written as

$$\log\left[\frac{P(y_i = 1)}{1 - P(y_i = 1)}\right] = x_i'\beta, \tag{3.7}$$

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$(\mu_i/(1 - \mu))$ which relates the expected value of the outcome variable to the linear predictor.

Similarly, the Poisson regression model Cameron and Trivedi (1998), which is used to model count data,is written as

$$\mu_i = exp(x_i'\beta) \tag{3.8}$$

or

$$log(\mu_i) = x_i'\beta, \tag{3.9}$$

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 covariates. This is done as

$$V(y_i) = \phi v(\mu_i), \tag{3.10}$$

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

$$v(\mu_i) = \mu_i(1 - \mu_i) \tag{3.11}$$

and ϕ is typically 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 mean equals the variance, and so

$$v(\mu_i) = \mu_i \tag{3.12}$$

and, again, ϕ is set to 1 (i.e., it is not estimated) in the usual GLM. As can be appreciated, the link function and variance specification usually depend on the distribution of the outcome variable y. Here we have presented three versions of GLMs but there are many others that are within the GLM family. With these GLM specifications, one can estimate the regression coefficients β by solving the estimating equation.

$$U(\beta) = \sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right)' (V(y_i))^{-1} [y_i - \mu_i] = 0 \tag{3.13}$$

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

$$U(\beta) = \sum_{i=1}^{N} x_i(y_i - x_i'\beta) = 0 \tag{3.14}$$

for solution of the regression coefficients β.

As noted by Wedderburn (1974), the above estimating equation (3.11) depends only on the mean and variance of y, and therefore the precise distribution form for y is not necessary for estimation of the regression coefficient β.In this case, solution of this estimating equation provides what are called 'quasi-likelihood' estimates.

GLMs are fixed effects models which assumes 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 longtitudinal data, and this is what Liang and Zeger did in developing GEE models.

3.2.2 Quasi-liklihood Estimates

The above GLM theory depends on choosing a distributional form for the data (eg. 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 $Q(y_i, \mu_i)$ is defined as:

$$\frac{\partial Q(y_i, \mu_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{V(\mu_i)} \tag{3.15}$$

or equivalently

$$\int^{\mu_i} \frac{y_i - \mu_i}{V(\mu_i)} d_i' + f(y_i). \tag{3.16}$$

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 overdispersed Poisson distribution. If the link and the variance correspond to a particular member of the exponential family, then the quasi-likelihood is equal to the likelihood proper.

3.3 GENERALIZED ESTIMATING EQUATIONS (GEE) MODELS

3.3.1 Overview

In the 1980s, alongside the development of Mixed-effect Regression Model (MRMs) and Covariance Pattern Models (CPMs) for incomplete longtitudinal data, generalized estimating equations (GEE) models were developed (Liang and Zeger (1986); Zeger and Liang (1986); Zeger et al. (1988)). Essentially, GEE model extend generalized linear models (GLMs) for the situation of correlated data. Thus this class of model 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 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 responds depends only on the covariates of interest and not on any random effects ~~or~~ previous responses". A basic premise of the GEE approach is that one is primarily interested in the regression parameters β and is not interested in the variance-covariance matrix of the repeated measures. As such the variance-covariance would be treated as a nuisance that one must account for in some way in order to get reasonable statistical tests for the regression parameters. Thus, the GEE models are not meant for situations in which scientific interest centers around the variance and/or covariance parameters.

With GEE the relationships between the variables of the model at different time-points are analysed simultaneously. So, the estimated β_1 reflects the relationship between the longitudinal development of the outcome variable Y and the longitudinal development of corresponding predictor variables X,using all available longitudinal data. GEE is an iterative procedure, using quasi-likelihood to estimate the regression coefficients (Liang and Zeger (1986); Zeger and Liang (1986); Zeger et al. (1988); Zeger and Liang (1992);Liang and Zeger (1993); Lipatz et al. (994a)).

A basic feature of GEE models is that the joint distribution of a subjects response vectory, 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 p, 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

$$\eta_{ij} = x'_{ij}\beta, \tag{3.17}$$

where x_{ij} is the covariate vector for subject i at time j . Then a link function

$$g(\mu_{ij}) = \eta_{ij} \tag{3.18}$$

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,

$$V(y_{ij}) = \phi v(\mu_{ij}), \tag{3.19}$$

where, again, $v(\mu_{ij})$ is a known variance function and ϕ is a scale parameter that may be known or estimated. 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 x n because one assumes that there are a fixed number of timepoints n that subjects are measured at. A given subject does not have to be measured at all n timepoints; each individuals correlation matrix $\mathbf{R_i}$ is of size n_i x n_i with the appropriate rows and columns removed if $n_i < n$. It is assumed that the correlation matrix $\mathbf{R_i}$, and thus $\mathbf{R_i}$, depends on a vector of association parameters denoted a. Examples of the various working correlation structures below will make this notion more concrete. These parameters a are assumed to be the same for all subjects. They represent the average dependence among the repeated observations across subjects. It is generally recommended that choice of **R** should be consistent with the observed correlations. However, the GEE method yields consistent estimates of the regression coefficients and their standard errors, even with misspecification of the correlational structure.This is a very attractive property of GEE models. Efficiency (i.e., statistical power) is reduced if the choice of **R** is incorrect; however, the loss of efficiency is lessened as the number of subjects gets large. The usual working correlations considered are independence,exhangeable, AR(1), m-dependent, and unspecified.

3.3.2 Working Correlation Structures

Because the repeated observations within one subject are not independent of each other, a correction must be made for these within-subject correlations. With GEE, this correction is carried out by assuming a priori a certain 'working correlation structure for the repeated measurements of the outcome ~~variable~~ Y . Depending on the software package used to estimate the regression coefficients, there is a choice between various correlation structures. The first possibility is an **independent structure**. With this structure the correlations between subsequent measurements are assumed to be zero.In fact, this option is counter-intuitive because a special technique is being used to correct for the dependency of the observations and this correlation structure assumes independence of the observations:

	t_1	t_2	t_3	t_4	t_5	t_6
t_1	1	0	0	0	0	0
t_2	0	1	0	0	0	0
t_3	0	0	1	0	0	0
t_4	0	0	0	1	0	0
t_5	0	0	0	0	1	0
t_6	0	0	0	0	0	1

A second possible choice for a working correlation structure is an **exchangeable structure**. In this structure the correlations between subsequent measurements are assumed to be the same, irrespective of the length of the time interval:

	t_1	t_2	t_3	t_4	t_5	t_6
t_1	1	ρ	ρ	ρ	ρ	ρ
t_2	ρ	1	ρ	ρ	ρ	ρ
t_3	ρ	ρ	1	ρ	ρ	ρ
t_4	ρ	ρ	ρ	1	ρ	ρ
t_5	ρ	ρ	ρ	ρ	1	ρ
t_6	ρ	ρ	ρ	ρ	ρ	1

A third possible working correlation structure, the so-called (**stationary**) m-**dependent structure** assumes that the correlations t measurements apart are equal, the correlations $t + 1$ measurements apart are assumed to be equal,and so on for $t = 1$ to $t = m$. Correlations more than m measurements apart are assumed to be zero. When, for instance, a 2-dependent correlation structure is assumed, all correlations one measurement apart are assumed to be the same, all correlations two measurements apart are assumed to be the same, and the correlations more than two measurements apart are assumed to be zero:

	t_1	t_2	t_3	t_4	t_5	t_6
t_1	1	ρ_1	ρ_2	0	0	0
t_2	ρ_1	1	ρ_1	ρ_2	0	0
t_3	ρ_2	ρ_1	1	ρ_1	ρ_2	0
t_4	0	ρ_2	ρ_1	1	ρ_1	ρ_2
t_5	0	0	ρ_2	ρ_1	1	ρ_1
t_6	0	0	0	ρ_2	ρ_1	1

A fourth possibility is an **autoregressive correlation structure**, i.e. the correlations one measurement apart are assumed to be ρ; correlations two measurements apart are assumed to be ρ^2 ; correlations t measurements apart are assumed to be ρ^t .

	t_1	t_2	t_3	t_4	t_5	t_6
t_1	1	ρ^1	ρ^2	ρ^3	ρ^4	ρ^5
t_2	ρ^1	1	ρ^1	ρ^2	ρ^3	ρ^4
t_3	ρ^2	ρ^1	1	ρ^1	ρ^2	ρ^3
t_4	ρ^3	ρ^2	ρ^1	1	ρ^1	ρ^2
t_5	ρ^4	ρ^3	ρ^2	ρ^1	1	ρ^1
t_6	ρ^5	ρ^4	ρ^3	ρ^2	ρ^1	1

The least restrictive correlation structure, is the **unstructured correlation structure**. With this structure, all correlations are assumed to be different:

$$
\begin{array}{c|cccccc}
 & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \\
\hline
t_1 & 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 & \rho_5 \\
t_2 & \rho_1 & 1 & \rho_6 & \rho_7 & \rho_8 & \rho_9 \\
t_3 & \rho_2 & \rho_6 & 1 & \rho_{10} & \rho_{11} & \rho_{12} \\
t_4 & \rho_3 & \rho_7 & \rho_{10} & 1 & \rho_{13} & \rho_{14} \\
t_5 & \rho_4 & \rho_8 & \rho_{11} & \rho_{13} & 1 & \rho_{15} \\
t_6 & \rho_5 & \rho_9 & \rho_{12} & \rho_{14} & \rho_{15} & 1
\end{array}
$$

It is assumed that GEE analysis is robust against a wrong choice of correlation matrix (i.e. it does not matter much which correlation structure is chosen, the results of the longitudinal analysis will be more or less the same) (Liang and Zeger (1986); Zeger and Liang (1986)). However, when the results of analyses with different working correlation structures are compared to each other, they differ in such a way that they can lead to wrong conclusions about longitudinal relationships between several variables (Twisk et al. (998a)). It is therefore important to realize which correlation structure is most appropriate for the analysis. Unfortunately, with GEE there is no straightforward way to determine which correlation structure should be used. One of the possibilities is to analyse the within-subject correlation structure of the observed data to find out which possible structure is the best approximation of the real correlation structure 1 . Furthermore, the simplicity of the correlation structure has to be taken into account when choosing a certain working correlation structure. The number of parameters (in this case correlation coefficients) that need to be estimated differs for each of the various working correlation structures. For instance, for an exchangeable structure only one correlation coefficient has to be estimated, while for a stationary 5-dependent structure, five correlation coefficients must be estimated. Assuming an unstructured correlation structure in a longitudinal study with six repeated measurements, 15 correlation coefficients must be estimated. As a result, the power of the statistical analysis is influenced by the choice of a certain structure. Basically, the best choice is the simplest correlation structure which fits the data well.In order to enhance insight in GEE analysis, the estimation procedure can be seen as follows. First a naive linear regression analysis is carried out, assuming the observations within subjects are independent. Then, based on the residuals of this analysis, the parameters of the working correlation matrix are calculated. The last step is to re-estimate the regression coefficients, correcting for the dependency of the observations. Although the whole procedure is slightly more complicated (i.e. the estimation process alternates between steps two and three, until the estimates of the regression coefficients and standard errors stabilize), it basically consists of the three above-mentioned steps (Burton et al. (1998)). In GEE analysis, the within-subject correlation structure is treated as a nuisance variable (i.e. as a covariate). So, in principle, the way in which GEE analysis corrects for the dependency of observations within one subject is the way that has been shown in Equation (3.20) (which can be seen as an extension of Equation (3.19)).

$$Y_{it} = \beta_0 + \sum_{j=1}^{J} \beta_{1j} X_{itj} + \beta_2 t + ... + CORR_{it} + \varepsilon_{it} \tag{3.20}$$

where Y_{it} are observations for subject i at time t, β_0 is the intercept, X_{ijt} is the independent variable j for subject i at time t, β_{1j} is the regression coefficient for independent variable j , J is the number of independent variables,t is time, β_2 is the regression coefficient for time, $CORR_{it}$ is the working correlation structure, and ε_{it} is the error for subject i at time t and it is normally distributed.

3.4 THE MATRIX EIGENVALUE PROBLEM

3.4.1 GEE Estimation

Define A_i to be the nxn diagonal matrix with $V(\mu_{ij})$ as the jth diagonal element. Also,as indicated above, define $R_i(a)$ be the nxn working correlation matrix (of the n repeated measures) for ~~the~~ i subject. ~~Then, the~~ working variance-covariance matrix for y_i equals

$$V(a) = \phi A_i^{1/2} R_i(a) A_i^{1/2} \tag{3.21}$$

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

$$V(a) = \phi R_i(a) \tag{3.22}$$

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

$$\sum_{i=1}^{N} D_i'[V(\hat{a})]^{-1}(y_i - \mu_i) = 0 \tag{3.23}$$

where $\hat{a}$ is a consistent estimate of a and $D_i = \partial\mu_i/\partial\beta$.Notice that this formula is an extension of the estimating equation for β in any GLM, which is given in (3.23). 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 (3.22)

$$\mu_i = X_i\beta,$$

$$D_i = X_i,$$

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

and so

$$\sum_{i=1}^{N} X_i'[R_i(\hat{a})]^{-1}(y_i - X_i\beta) = 0 \tag{3.24}$$

which yields, solving for β

$$\hat{\beta} = \left[\sum_{i=1}^{N} X_i'[R_i(\hat{a})]^{-1}X_i\right]^{-1}\left[\sum_{i=1}^{N} X_i'[R_i(\hat{a})]^{-1}y_i\right] \tag{3.25}$$

Notice this is essentially the same as the Weighted Least-Squares (WLS) estimator with the weight matrix being $[R_i(\hat{a})]^{-1}$. A caveat is that in WLS one typically knows the weight matrix, however here it depends on parameters to be estimated (i.e,a). In this case, the solution can proceed using Iteratively Reweighted Least Squares (IRLS) where iterative estimates of a are used to yield new estimates of β, with the procedure continuing until convergence. Because equation (3,23) depends only on the mean and variance of y, these are quasi-likelihood estimates.

Solving the GEE involves iterating between the quasi-likelihood solution for estimating β and a robust method for estimating a as a function of β. Essentially, it involves the following steps, which are repeated until convergence.

1. Given estimates of $R_i(a)$ and ϕ, calculate estimates of β using IRLS.

2. Given estimates of β, obtain estimates of a and ϕ. For this, calculate Pearson (or standardized) residuals

$$r_{ij} = (y_{ij} - \hat{\mu_{ij}})/\sqrt{[V(\hat{a})]_{jj}} \tag{3.26}$$

and use these residuals to consistently estimate a and ϕ. Liang and Zeger (1986) 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 $\hat{\beta}$

$$V(\hat{\beta}) = \left[\sum_{i}^{N} D_i' \hat{V}_i^{-1} D_i\right]^{-1} \tag{3.27}$$

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 Estimator

The estimator

$$V(\hat{\beta}) = M_0^{-1} M_1 M_0^{-1} \tag{3.28}$$

represents the covariance matrix of $\hat{\beta}$. It has the property of being a consistent estimator matrix of $\hat{\beta}$ even if the working correlation matrix is misspecified

where

$$M_0 = \sum_i^N D_i' \hat{V}_i^{-1} D_i \tag{3.29}$$

$$M_1 = \sum_i^N D_i' \hat{V}_i^{-1} (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)' \hat{V}_i^{-1} D_i \tag{3.30}$$

Here, $\hat{V}_i$ denotes $V_i(\hat{a})$. 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. In the more general case, the robust or sandwich estimator, which is due to Royall (1986), provides a consistent estimator of $V(\hat{\beta})$ even if the working correlation structure $\mathbf{R_i(a)}$ is not the true correlation of y_i.

3.5 RANDOM EFFECT ANALYSIS

3.5.1 Overview

Random effects analysis is also known as multilevel analysis or random coefficient analysis Laird and Ware (1986); Longford (1993); Goldstein (1991). Multilevel analysis was initially developed in the social sciences, more specifically for educational research. This type of study design is characterized by a hierarchical structure. Students are nested within classes, and classes are nested within schools. Various levels can be distinguished. As this technique is suitable for correlated observations, it is obvious that it is also suitable for use in longitudinal studies. In longitudinal studies the observations within one subject over time are correlated. The observations over time are nested within the subject. The basic idea behind the use of multilevel techniques in longitudinal studies is that the regression coefficients are allowed to differ between subjects. Therefore the term random coefficient analysis is preferred to the term multilevel analysis.

3.5.2 Random Effects Analysis In Longitudinal Studies

The simplest form of random effects analysis is an analysis with only a random intercept. The corresponding statistical model with which to analyse a longitudinal relationship between an outcome variable Y and time is given in Equation (3.31).

$$Y_{it} = \beta_{0i} + \beta_1 t + \varepsilon_{it} \tag{3.31}$$

where Y_{it} are observations for subject i at time t, β_{0i} is the random intercept,t is time, β_1 is the regression coefficient for time, and ε_{it} is the error for subject i at time t. What is new about this model (compared to Equation (3.2)) is the random intercept β_{0i} , i.e. the intercept can vary between subjects. It is also possible that the intercept is not random, but that the development of a certain variable over time is allowed to vary among subjects or, in other words, the slope with time is considered to be random.

$$Y_{it} = \beta_0 + \beta_{1i}t + \varepsilon_{it} \tag{3.32}$$

where Y_{it} are observations for subject i at time t, β_0 is the intercept, t is time, β_{1i} is the random regression coefficient for time, and ε_{it} is the error for subject i at time t. The most interesting possibility is the combination of a random intercept and a random slope with time, which is in Equation (below).

$$Y_{it} = \beta_{0i} + \beta_{1i}t + \varepsilon_{it} \tag{3.33}$$

where Y_{it} are observations for subject i at time t, β_{0i} is the random intercept, t is time, β_{1i} is the random regression coefficient for time, and ε_{it} is the error for subject i at time t.

The assumption of random coefficient analysis is that the variation in intercept and variation in slopes are normally distributed with an average of zero and a certain variance. This variance is estimated by the statistical software package. The general idea of random coefficient analysis is that the unexplained variance in outcome variable Y is divided into different components. One of the components is related to the random intercept and another component is related to random slopes.

For longitudinal studies, random effects models enable the analyst to not only describe the trend over time while taking into account of the correlation that exist between massive measurements, but also to describe the variation in the baseline measurement and in the rate of change over time.

3.5.3 Nature Of Data Collection

- Subjects are not assumed to be measured on the same number of time points, and the time points do not need to be necessarily equally spaced.
- Analyses can be conducted for subjects who may miss one or more of the measurement occasions, or who may be lost to follow-up at some point during study. In our example a student may fall sick during the entire duration of a semester examination.

Random effects models however allow for the inclusion of time-varying and time-invariant covariates. Time-varying covariates are independent variables that co-vary with the dependent variable over time. For example, a researcher studying trends in students Y_i performance over time, might also want to capture data on the highest and lowest marks of the group or degree of performance of co-student at each measurement occasion. The background of the student is likely to be an important predictor for performance assessment which may also vary over time. Covariates such as gender and geographical location status either do not change over time, or are less likely to change over time.

Random effects allow the research analyst to model the correlation structure of the data. Thus, the analyst does not need to assume that measurements taken at successive times in time are equally correlated, which is the correlation structure that underlies the ANOVA model. The analyst also does not need to assume measurements taken at successive points in time have an unstructured pattern of correlations, which is the structure that underlies the multivariate analysis of variance model. The former pattern is generally too restrictive, while the latter is too generic. With random effects model, analyst can fit a specific correlation structure to the data, such as an autoregressive structure, which assumes a decreasing correlation between successive measurements over time.

3.5.4 Random Intercept Model(REM)

The simplest regression model for longitudinal data is one in which measurements are obtained for a single dependent varaiable at successive time points. Let Y_{ij} represent the measurement for the ith individual at the jth point in time,

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij} \tag{3.34}$$

β_0 is the intercept, β_1 is the slope, that is the change in the outcome variable for every one-unit increase in time(semester) and ε_{ij} is the error component. In this simple regression the ε_{ij}'s are assumed to be correlated and to follow a normal distribution (ie. $\varepsilon_{ij} \sim N(0, \sigma^2)$). β_0 represents the average value of the dependent variable when time=0 and β_1 represents the average change of the dependent variable for each one-unit increase in time(semester). There is a possibility that a student may start with a low SWA and then increase over semesters as shown in Figure 3.1 or no change in SWA over semesters as shown in Figure

3.2 or start with a high SWA and decrease over semester as in Figure 3.3

The implication was that on average students who perform well is depicted by Figure 3.1 whereas Figure 3.3 depicts that on the average students are performing poorly.

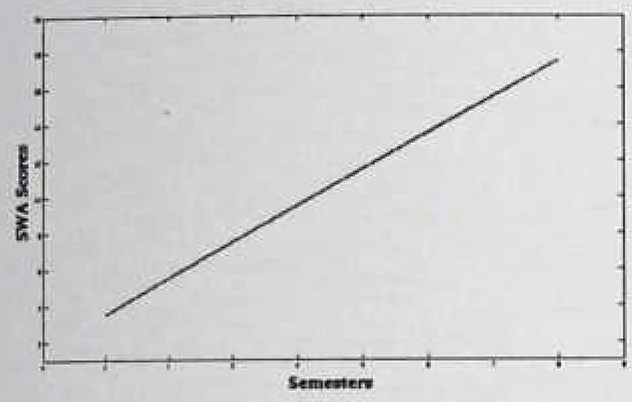

Figure 3.1: Positive slope

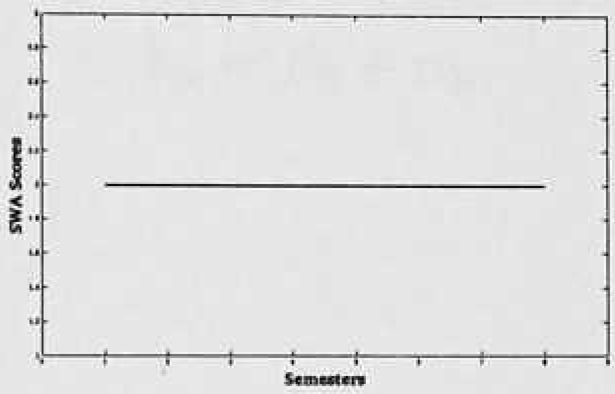

Figure 3.2: Constant slope

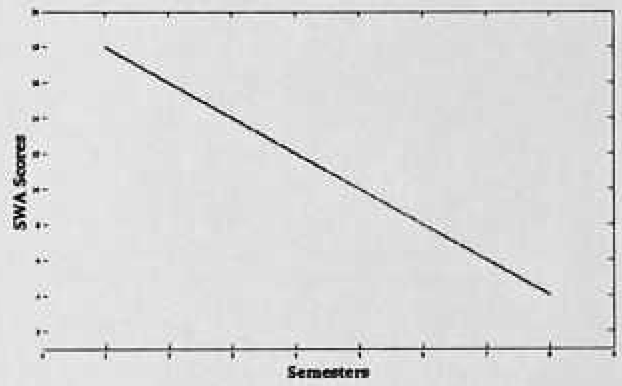

Figure 3.3: Negative slope

Figures 3.1, 3.2 & 3.3 : Possible Average Change in SWA's over Semesters

The simple random effect is the one which the intercept is allowed to vary across individuals (students):

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \upsilon_{0i} + \varepsilon_{ij} \tag{3.35}$$

where υ_{0i} represent the influence on individual i on his/her repeated observations. Note that is the individuals have no influence on their repeated outcome(SWA), then all the υ_{0i} will be equal to zero (υ_{0i}=0), but that may not be true. Therefore υ_{0i} may have negative or positive impact on their SWA's therefore υ_{0i} may deviate from zero. for better reflection of this model on the characteristic individual, the model is partition into within-subjects and between-subjects.

Within-subject

$$Y_{ij} = b_{oi} + b_1 t_{ij} + \varepsilon_{ij} \tag{3.36}$$

Between subjects

$$b_{oi} = \beta_0 + \upsilon_{0i} \tag{3.37}$$

$$b_{1i} = \beta_1 \tag{3.38}$$

Equation (second above) indicates that the intercept for the ith individual is a function of a population intercept plus unique contribution for individual. We assume $\upsilon_{0i} \sim N(0, \sigma^2)$. This model also indicates that each individual's slope is equal to the population slope, β_1, equation(last one above)

When both the slope and the intercept are allowed to vary across individual the model is:

$$Y_i = \beta_0 + \beta_1 t_{ij} + \upsilon_{0i} + \upsilon_{1i} t_{ij} + \varepsilon_i \tag{3.39}$$

The within-subjects model is the same as

$$Y_i = b_{oi} + b_{1i} t_{ij} + \varepsilon_i \tag{3.40}$$

The between-subjects models is:

$$b_{oi} = \beta_0 + v_{0i} \tag{3.41}$$

$$b_{1i} = \beta_1 + v_{1i} \tag{3.42}$$

The within-subject model indicates that the individual ith SWA at time j is influenced by their initial level b_{0i} and the time trend or the slope b_{1i}. The between-subject indicates that the individual's i's initial level is determined by the population initial level β_0 plus the unique contribution of v_{0i}. Thus each individual has their own distinct initial level. Intercept for the ith individual is a function of a population intercept plus unique contribution for that indivual. As well, the slope for the ith individual is a function of the population slope plus some unique contribution for that subject. We assume

is the variance-covariance matrix of the random effects. Correlation exists between the random slope and the random intercept, so that individuals who have higher values for the intercept (ie. higher or lower ~~values for~~ the slope). The resulting linear model can now be written as:

$$Y_i = X_i\beta + Z_ib_i + \varepsilon_{1i} \tag{3.43}$$

$$b_i \sim N(0, D)$$

$$\varepsilon_{1i} \sim N(0, \sigma^2 I_{ni})$$

Assumptions:

- $b_1, \ldots, b_N, \varepsilon_1, \ldots, \varepsilon_N$ b's are independent

- $\varepsilon_1 \sim N(0, \sigma^2 I_{ni})$ is the measurement error

The variance of the measurement is given below

$$V(y_i) = Z_i \sum_{v} Z_i^{'} + \sigma^2 I_{ni} \tag{3.44}$$

This model implies that conditional on the random effects, the errors are uncorrelated, as is displayed. This is seen in the above equation (eqn above) since the error variance is multiplied by the identity matrix (i.e., all correlations of the error equal zero).

3.5.5 Restricted Maximum Likelihood Estimation

Here we consider the case where the variance of a normal distribution $N(\mu, \sigma^2)i$ to be estimated based on a sample $Y_1, \ldots, Y_N$ of N observations. Where the mean μ is known, the maximum likelihood estimation (MLE) for σ^2 equals $\widehat{\sigma^2} = \sum_i \frac{(Y_i - \mu)^2}{N}$ which is unbiased for σ^2 . When μ is not unknown, we get the same expression for the MLE but with μ

replaced by the sample mean $\overline{Y} = \sum_i \frac{Y_i}{N}$

$$E(\widehat{\sigma^2}) = \frac{N-1}{N}\sigma^2 \qquad (3.45)$$

The equation (as in above) as in above indicate that the MLE is now biased downward due to the estimation of μ, the unbiased estimation of (as in above) yield the classical samle variance.

$$S^2 = \sum_i \frac{(Y_i - \overline{Y})^2}{N-1} \qquad (3.46)$$

To obtain an unbiased estimate for σ^2 directly we should use the following: let $Y = (Y_1, ..., Y_N$ denote the vector for all measurement and I_N be $N - \textit{dimensional}$ vector containing only ones and zeros. The distribution of Y is then $N(\mu I_N, \sigma^2 I_N)$ where I_N equals the identity matrix, if A is $N \times (N-1)$ any matrix with $N-1$ linear independent columns orthogonal to the vector I_N, vector U of $N-1$ which is the error contrast is defined by $U = A^T A$. Maximizing the corresponding likelihood with respect to the only remaining parameter σ^2 yields $\widehat{\sigma^2} = \frac{Y_T A(A^T A)^{-1} A^T Y}{N-1}$ which is equal to classical sample variance S^2. The resulting estimator is the RMLE since it rectricts $(N-1)$ error contrasts.

3.5.6 The Random Intercepts Model

The random effects covariance matrix D is now scalar and it will be denoted by σ_b^2, the matrix Z_i are of the form I_{n_i}, a $n_i - \textit{dimensional}$ vector of ones. We will assume that all residual covariance matrices are of the form $\sum_i = \sigma^2 I_{n_i}$, i.e., we assume conditional independent. The random intercept of subject i is given by

$$\widehat{b_i} = \sigma_b^2 I'_{n_i}(\sigma_b^2 I_{n_i} I'_{n_i})^{-1}(y_i - X_i\beta) \tag{3.47}$$

$$= \sigma_b^2 I'_{n_i}(\frac{\sigma_b^2}{\sigma^2} - \frac{\sigma_b^2}{\sigma^2 + n_i\sigma_b^2} I_{n_i} I'_{n_i})(y_i - X_i\beta) \tag{3.48}$$

$$= \frac{n_i\sigma_b^2}{\sigma^2 + n_i\sigma_b^2}\frac{1}{n_i}\sum_{j=1}^{n_i}(y_i - x'_{ij}\beta) \tag{3.49}$$

Where the vector x_{ij} consists of the jth row in the design matrix X_i and $\frac{1}{n_i}\sum_{j=1}^{n_i}(y_i - x'_{ij}\beta)$ is equal to the average residual. If n_i is large for subject i a weight is put on the average residual yielding less shrinkage, the within-subject variability is large in comparison to the between subject variability if more shrinkage is obtained.

Chapter 4

SUMMARY OF RESULTS AND DISCUSSION

4.1 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 aggregate, geographical location of students as well as the Ghana Education Service Categories of Schools of former SHS attended. The chapter however seeks to model the students' academic performance based on Semester Weighted Average (SWA) using the Generalized Estimating Equation model (GEE) and randon effect with respect to the working correlation assumptions.

4.1.1 Data Collection

The consecutive students' Semester Weighted Average (SWA) academic results of Class of 2012 Mathematics students at KNUST (i.e. eight semesters) were obtained. The obtained SWA(s) and their socio-demographic factor response variables were tallied and also coded in the Window Microsoft Excel (2010), R and the SAS version 9.1 for the analyses. The geographical locations of students were categorized into three zonal belts specifying their respective regions of origin. These include the Northern Belt (includes Northern, Upper East, Upper West) coded as N, Middle Belt (Ashanti, Brong Ahafo Regions, Eastern, Volta) was coded M and Southern Belt (includes Greater Accra, Central and Western regions) were also coded as S. Similarly, the graded schools were categorized into A, B, C, D and P from Ghana Education Service (GES) specification. Grade A schools were coded as A, grade B schools were coded as B,grade C schools were coded as C, grade D schools were coded as D and grade P schools were coded as P.There were ninety (90) students in the class of 2012 Mathematics students records sampled. Details of the analyses are discussed below.

4.1.2 Exploratory Data Analysis

This subsection of the chapter highlights on the descriptive data analyses for the study with respect to the various research variables. Descriptive statistics are used to describe the basic features as well as pictorial view of the data gathered from the study. They provide summaries of the sample and the measures. Together with simple graphic analysis, they form the basis of virtually every quantitative analysis. This means that being the location parameter of the distributions, tell little about the data when left in isolation.The mean and the correlation structure were also considered for the data sets.Correlation structure describes how SWA obtained are correlated over the Semesters. The correlation function depends on a pair of SWA scores over the semesters.The mean structure, which is the average evolution describes how the profile for students with respect to SWA on the average evolves over the semesters conditioned on the type of school, entry aggregate, geographical location etc. The results of this exploration will be useful in order to choose a fixed-effect structure for the random effect model. 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 means. The smaller the standard deviation, the closer the data points to the mean whereas the larger the standard deviation, the less representative would be the mean. The variance structure is also important to build an appropriate longitudinal model. Here it is necessary that we correct the measurements for the fixed effects structure.

Table 4.1: Gender of Students in Maths Four

Gender	Frequency (f)	Percent (%)	Cummulative f	Cummulative %
F	15	16.85	15	16.85
M	74	83.15	89	100.00

Frequency missing=1

We observe from table 4.1 that, out of a total of 90 students, 15 students coded F were females representing 16.85% and 74 students coded M were males representing 83.15%.The gender of a student was not mentioned in the data and therefore was treated as missing.

Table 4.2: Geographical Location of Students in Maths Four

Geographical Location	Frequency (f)	Percent (%)	Cummulative f	Cummulative %
M	50	58.14	50	58.14
N	10	11.63	60	69.77
S	26	30.23	86	100.00

Frequency missing=4

The Geographical Location of Maths Four students as in Table 4.2 was grouped into three zonal belts specifying their respective regions of origin. These include the Northern Belt (includes Northern, Upper East, Upper West) coded as N, Middle Belt (Ashanti, Brong Ahafo Regions, Eastern, Volta) was coded M and Southern/Coastal Belt (includes Greater Accra, Central and Western regions) were also coded as S have been specified. Fifty (50) of the students representing 58.14% represented the Middle Belt was coded as M. Also, ten (10) of the students representing 11.63% represented the Northern Belt was coded as N. Similarly, twenty-six (26) of the students representing 30.23% represented the Southern Belt was coded as S. The Geographical Location of four (4) students was not captured in the data and therefore was treated as missing.

From Table 4.3, Maths Four students are grouped according to the GES special SHS categorization with respect to the type of SHS attended; whether Grade A, B, C, D or P

Table 4.3: Ghana Education Service Categories of Schools in Maths Four

GES Categories	Frequency (f)	Percent (%)	Cummulative f	Cummulative %
A	47	55.29	47	55.29
B	25	29.41	72	84.71
C	9	10.59	81	95.29
D	2	2.35	83	97.65
P	2	2.35	85	100.00

Frequency missing=5

schools are specified in the table. It could be observed that out of the total of 90 students, 47 students (representing 55.29%) attended a Grade A schools, 25 students (representing 29.41%) attended Grade B school, 9 students (representing 10.59%) attended a Grade C school, 2 students (representing 2.35%) attended both a Grade D school and a Grade P school respectively, whilst 5 students information failed to be captured and therefore was recorded as missing.

From the Table 4.4 the average SWA by the GES category over the semesters statistically generally increases from semester 1 through to semester 8. In semester 6 students from the Type C school recorded a mean SWA score with a value of 81.333%; incidentally is the highest SWA score recorded throughout the whole 8 semesters. However, in semester 4 students from the Type C school also recorded the least among the Types of schools as well as the least amongst all the 8 semesters under consideration with a mean SWA score of 54.758%. Students from Type B school achieved on the average the maximum mean

Table 4.4: Average GES Categories Statistics for Students in Maths Four

Semester	**Type A**	**Type B**	**Type C**
1	59.862	60.652	56.241
2	57.450	58.776	55.801
3	58.987	60.961	55.918
4	58.653	59.9818	54.758
5	60.926	60.583	58.389
6	62.902	62.167	81.333
7	64.074	64.578	65.569
8	68.899	68.216	66.825

SWA with a value of 60.652%. However, in semester 8, students from the Type A ended the program on a strong note and recorded a mean SWA of 68.899%;coincidentally the maximum mean SWA score throughout the 8 semesters.

It is observed from Table 4.5 that students SWA generally increased from semester 1 through to semester 8. Students from the Northern Belt had the highest average SWA in semester 1 with an approximate average SWA of 59.95 with students from the Middle Belt securing the least average SWA in semester 1. However in semester 8, on the average, students from the Middle Belt had the highest SWA relative to that of the Northern and Southern Belt with an appropriate average SWA of 68.90, 68.22 and 66.83 respectively.

We observe from Table 4.6 that the maximum SWA score over all the semesters is in the

Table 4.5: Average Geographical Location Statistics for Students in Maths Four

Semester	Middle Belt(M)	Northern Belt(N)	Southern Belt(S)
1	58.439	59.948	58.844
2	57.042	58.545	55.832
3	58.658	59.986	57.383
4	59.928	59.517	55.879
5	61.767	61.497	57.571
6	63.715	62.803	60.352
7	64.074	64.674	61.795
8	68.899	68.216	66.825

Table 4.6: Summary Statistics for the SWA's

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
SWA1	90	59.289	7.390	5336	44.222	75.833
SWA2	90	57.451	5.567	5171	44.211	68.842
SWA3	88	58.925	7.345	5185	41.421	76.053
SWA4	87	58.523	8.749	5091	36.864	76.632
SWA5	86	60.275	9.307	5184	30.833	82.833
SWA6	84	62.182	8.148	5223	37.000	81.333
SWA7	84	64.133	9.173	5387	40.294	81.600
SWA8	83	68.322	7.390	5671	49.952	86.952

fifth semester with a score of approximately 82.83% and the minimum SWA score over all the semesters is in fourth semester with a score of 36.86%. In a semester where a student was either deffered, repeated or dimissed recoreded a minmum percentage of zero (0).

It can also be inferred that the mean SWA of the 90 students over the eight(8) semesters were almost similar (i.e. around the average of 60% over the semesters). A small standard deviation depicts how near the data points are to the actual SWA mean.The semester with the highest variation of SWA is semester six (6)

Table 4.7: Variance - Covariance of SWAs for Maths Four Students

SWA	SWA1	SWA2	SWA3	SWA4	SWA5	SWA6	SWA7	SWA8
SWA1	54.615	32.325	42.711	47.742	47.989	36.404	38.078	30.949
SWA2	32.325	30.997	31.485	36.707	35.125	26.813	28.889	21.326
SWA3	42.711	31.485	53.948	52.959	55.294	42.665	46.987	35.188
SWA4	47.742	36.707	52.959	76.548	68.050	55.900	64.147	50.587
SWA5	47.989	35.125	55.294	68.050	86.613	65.611	71.377	55.400
SWA6	36.404	26.813	42.665	55.900	65.611	66.385	63.332	50.755
SWA7	38.078	28.889	46.987	64.147	71.377	63.332	84.144	59.169
SWA8	30.949	21.323	35.188	50.587	55.400	50.755	59.169	54.612

Table 4.7 statistically shows the measure of students covariances and variances of their SWAs over the semesters. Covariance is the key measure of the strength of relationship between the various semesters.

However in order to translate it into a measure one can understand we consider the correlation coefficient and for that matter the Pearson's Correlation Coefficient ($\hat{\rho}$).

Table 4.8: (8 x 8) Estimated Pearson Correlation Coefficients($\hat{\rho}$) Matrix For SWA's

Pearson Correlation Coefficients($\hat{\rho}$)								
SWA	**SWA1**	**SWA2**	**SWA3**	**SWA4**	**SWA5**	**SWA6**	**SWA7**	**SWA8**
SWA1	1.000	0.786	0.785	0.733	0.701	0.614	0.571	0.579
SWA2	0.786	1.000	0.778	0.758	0.680	0.603	0.577	0.535
SWA3	0.785	0.778	1.000	0.825	0.805	0.727	0.711	0.668
SWA4	0.733	0.758	0.825	1.000	0.831	0.791	0.806	0.786
SWA5	0.700	0.680	0.805	0.831	1.000	0.864	0.835	0.809
SWA6	0.614	0.603	0.727	0.791	0.864	1.000	0.847	0.840
SWA7	0.571	0.577	0.711	0.806	0.835	0.847	1.000	0.883
SWA8	0.579	0.535	0.668	0.786	0.809	0.840	0.883	1.000

It is observed from Table 4.8 that, statistically, students' SWAs are positively correlated. Generally, there is a slight decrease in the estimated pairwise Pearson Correlation Coefficients over the semesters. It is also observed that all the Pearson Correlation Coefficients are strong (> 0.5) and significant at 5%.However, the coefficient is stronger between adjacent semesters.

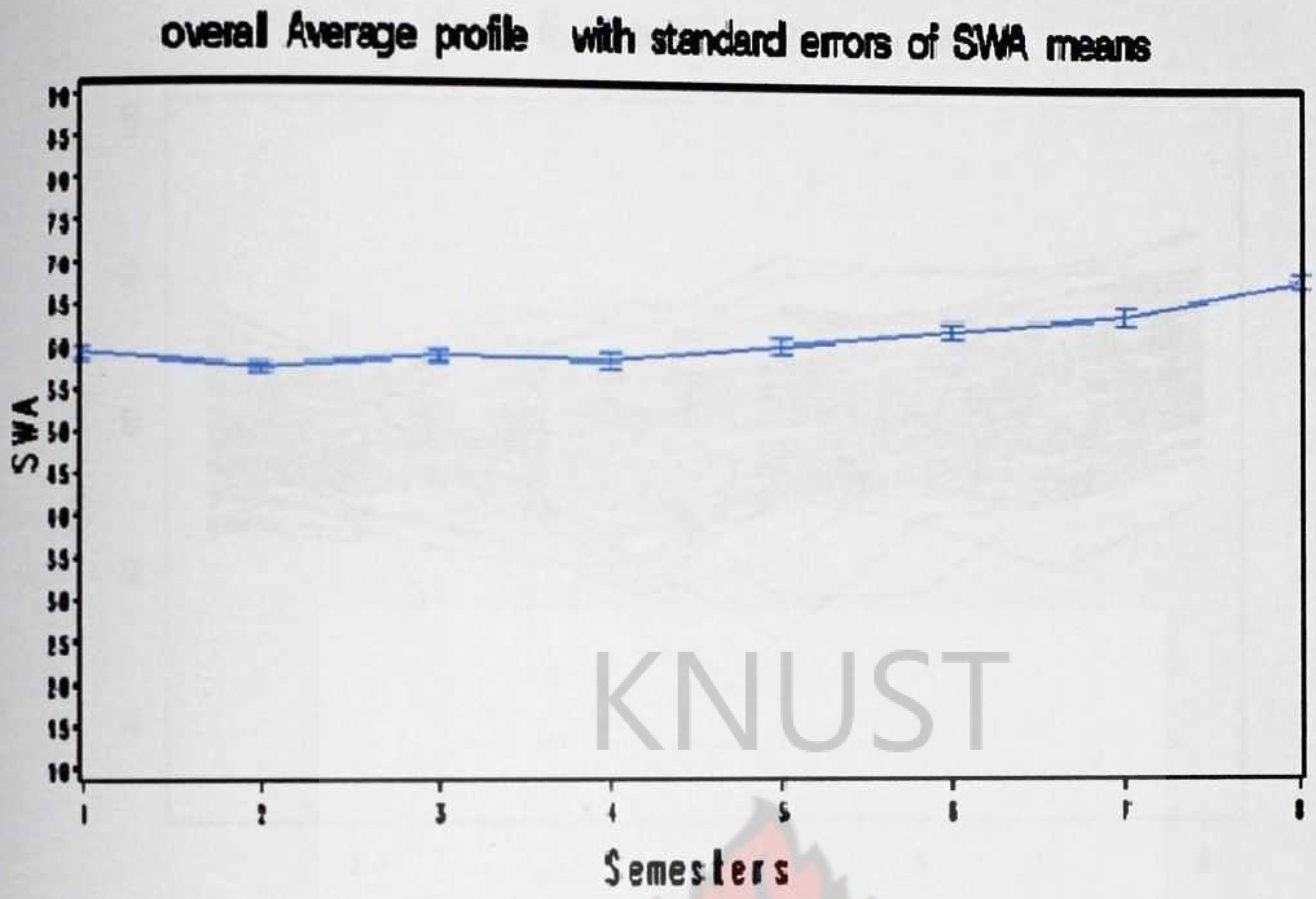

Figure 4.1: Overall Average Profile & Standard Errors with SWA Score

Figure 4.1 represents the overall average profile and standard errors with SWA score. Here it is observed that on the average all the students in Maths Four started with an average SWA score of 60.00%. The average performance of student in relation to the average SWA throughout the eight (8) semesters marginally increased as the semester progressed with minimal standard deviations.

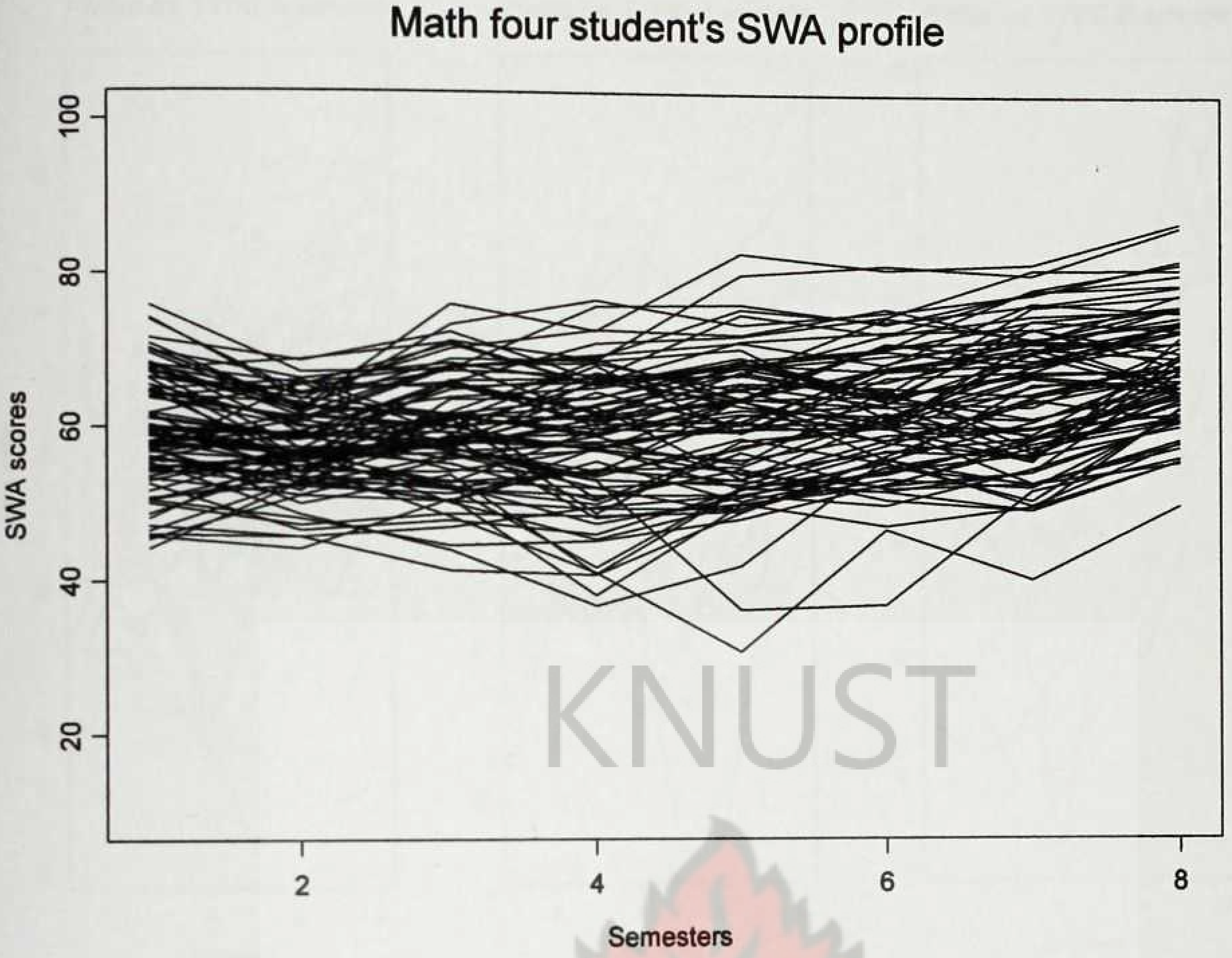

Figure 4.2: Maths Four (4) Students Profile

In general most students started with a low SWA but managed to complete with a higher SWA as in Figure 4.2. Students SWAs over the semesters lie between the 50 and the 70 mark; with semesters 8 and 5 approximately recording the highest and the least SWA of 86.9% and 30.8% respectively.

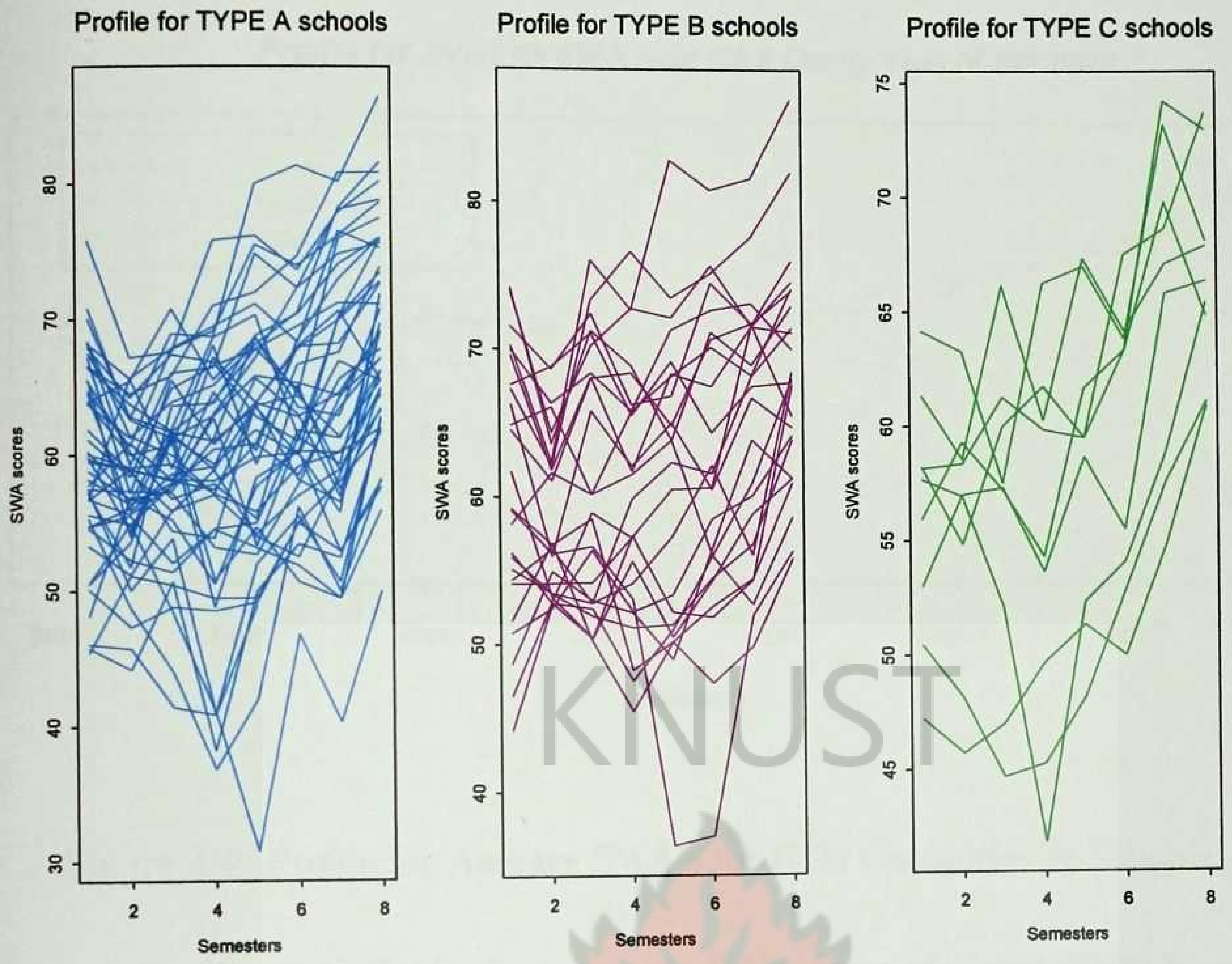

Figure 4.3: GES Categories of Schools in Maths Four

Most of the students in Maths Four were from a Type A school with a Type C school having the least number of students as shown in Figure 4.3. A student in Type B school recorded the highest SWA score of 86.9% in semester 8. However, a student from a Type A school recorded the least SWA score over the 8 semesters.

The mean profiles in Figure 4.4 are the means of the students SWA conditional for the various GES SHS gradings. It is observed that, on the average SWA's of students from the Type A school (about 69.4%) is higher relative to Type B and C over the semesters. About 80% of the all the students' average SWA lies between 55% and 65.8%. The average SWA's

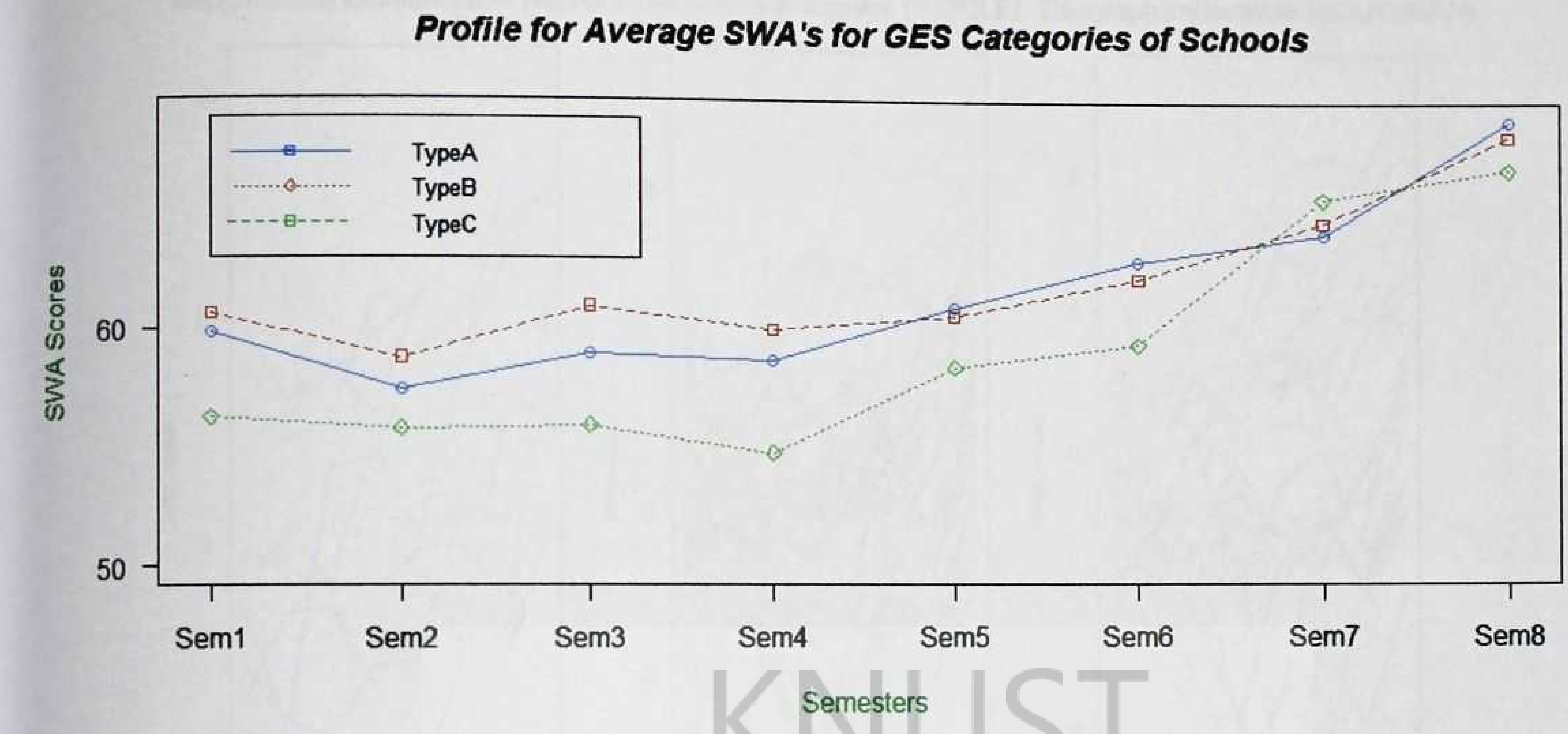

Figure 4.4: Profile for Average SWA's for GES Categories of Schools

of students in the Type A obtained an average SWA score of about 60.00 in semester 1 and later dropped in semester 2, the average performance droped slightly. However from semester 2 to semester 8, there was a gradual increase in the average SWA score. The averagė SWA's score obtained by students in the Type B school in semester 1 was about 60.00. The average SWA rose from semester 2 to 3 and slightly dropped from semester 3 to 4. There was also a consistent increase in the average SWA from semester 4 through to semester 8 where their highest average SWA score was recorded. The average SWA obtained by students from the Type C school have an average SWA score of about 57.6. The average performance was fairly stable from the semester 1 through to semester 3. However, there was a methodical and a consistent rise in the average SWA from semester 4 through to semester 8 where they had their highest average SWA and thereby catching

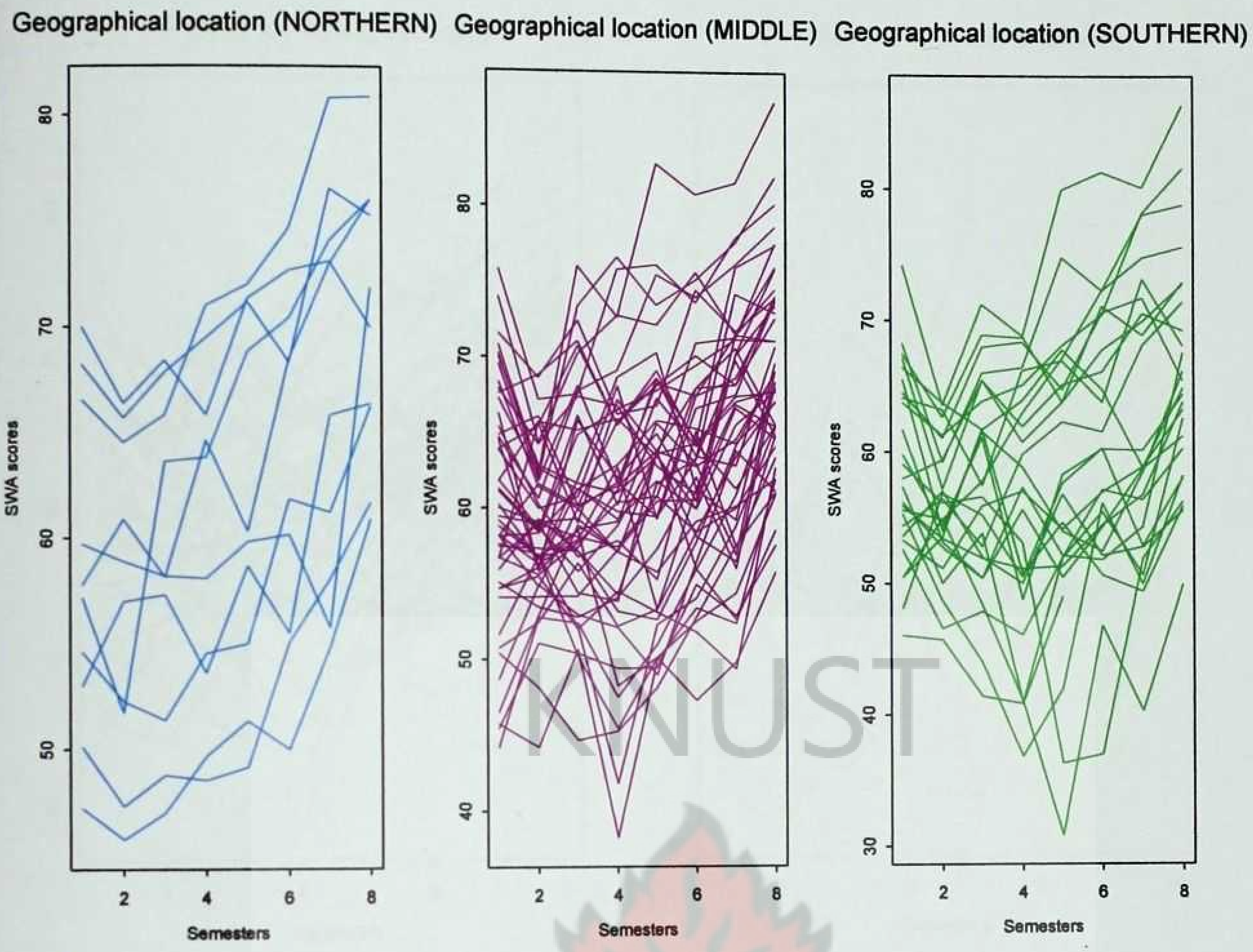

Figure 4.5: Geographical Location For Maths Four Students

up with the average performance with the Type A and B schools from semester 7 through to semester 8.

In general, majority of the students from Maths Four as in Figure 4.5 reside in the Middle Belt of Ghana whereas those ~~from the~~ Northern Belt contributed the least. The minimum SWA score was recorded by a student from the Southern Belt whereas the maximum SWA score was recorded by a student from the Middle Belt. The distribution of the SWAs of students from the Middle Belt and Southern was very dense approximately between the 51 - 68 % and 51 - 65 % respectively. However, the distribution of students due to their

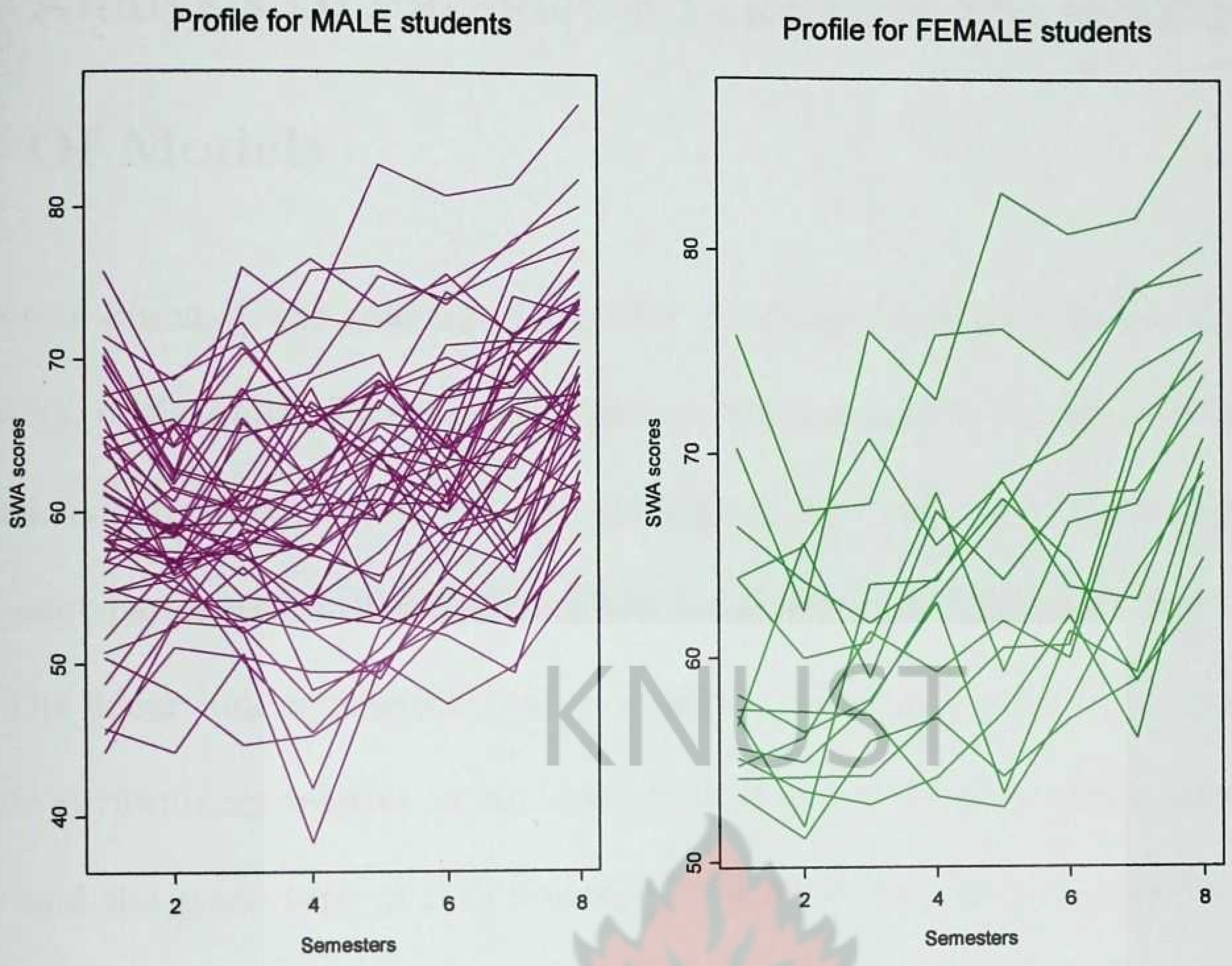

Figure 4.6: Gender Profile for Maths(4) Students

population was scattered over the 8 semesters around the 50 - 72 mark.

The profile of the male students in Maths Four as in Figure 4.6 is denser than that of the females indicating a greater number of males than females. Majority of the males SWAs lies approximately between 52 and 66 whereas that of the females lies between 55 and 65. Statistically, the males recorded a maximum and a minimum SWA of 84.2 and 36.8 whereas the females recorded a maximum and a minimum SWA of 86.9 and 51.8 over the 8 semesters.

4.2 Analysis Of Parameter Estimates For GEE Family Of Models

This section discusses the analyses of the GEE parameter estimates. In getting the right parameters, critical analyses are made based on working correlation assumptions already discussed. A main effect model and a model indicating a linear time interactions with the independent parameters are analysed. These are presented in Tables 4.8 and 4.9 respectively. The linear time interaction gives us a general implicative effect of time on students academic performance relative to students students geographical location, entry SHS aggregate and the grade type of SHS attended. The effect may seem to vary by treatment group. The subsections discusses the working correlation matrix for the four GEE models. It is generally advisable to choose a working correlation structure that is similar to the structure of the observed correlations.

Table 4.9 summarizes the computations for parameter estimates for our four (4) GEE working correlation assumptions already discussed. The parameter estimates for the four models are approximately the same for all the assumptions. All computations are approximated to three (3) decimal places. The two (2) standard errors (*Std Err*) for each assumption are stated: the *empirical based* and the *model-based* model. The standard error that is bracketed (.) represents the model based standard error. Gender F and M represents females and males, Region M, N and S represents the Middle, Northern and Southern belts, Entry Agg. represents Entry Aggregates whereas GES. CAT A, B, C, P represents GES SHS

Table 4.9: Estimated Coefficients and Standard Errors for GEE Main Effect Models

	INDEPENDENCE		AR(1)		UNSTRUCTURED		EXCHANGEABLE	
PARAMETERS	Est	Std Err	Est	Std Err	Est	Std Err	Est	Std Err
INTERCEPT	88.738	10.347(4.949)	90.217	10.039(10.439)	23.384	43.089(2.845)	90.146	10.567(10.657)
AGE	-0.529	0.277(0.166)	-0.584	0.276(0.349)	0.901	1.147(0.096)	-0.586	0.294(0.355)
GENDER(F)	3.420	1.877(0.867)	3.303	1.794(1.838)	0.954	9.883(0.954)	3.497	1.893(1.884)
GENDER(M)	0.000	0.000(0.000)	0.000	0.000(0.000)	0.000	0.000(0.000)	0.000	0.000(0.000)
REGION (M	3.030	1.881(0.759)	2.701	1.785(1.607)	-4.725	8.742(0.431)	3.169	1.887(1.645)
REGION (N)	7.729	2.484(1.429)	7.136	2.455(3.016)	-31.899	10.093(0.820)	8.119	2.567(3.080)
REGION(S)	0.000	0.000(0.000)	0.000	0.000(0.000)	0.000	0.000(0.000)	0.000	0.000(0.000)
ENTRY.AGG	-1.299	-1.299(0.190)	-1.325	0.457(0.402)	0.023	2.257(0.107)	-1.316	0.478(0.412)
GES.CAT A	-1.258	-1.258(2.102)	0.358	3.625(4.457)	25.017	9.400(1.187)	-1.490	3.913(4.566)
GES.CAT B	0.379	0.379(2.146)	2.001	3.514(4.553)	27.189	9.274(1.209)	0.275	3.855(4.667)
GES.CAT C	0.591	0.591(2.361)	1.893	3.825(5.01)	-14.142	13.134(1.329)	0.585	4.162(5.136)
GES.CAT P	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 model based Standard Error (Std Err)

category A, B, C, P respectively.

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. Therefore, the dependent variables are uncorrelated and are independent of each other across time (with regards to the eight semesters). Also perfect correlation is realized within and between the dependent variables under study and zero elsewhere. It could be inferred from Table 4.9 that the parameter estimates for the independent variables (age, gender, geographical locations and type of SHS school attended) are the same for both the empirical (robust) and model-based (*naïve*) 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 modelled using the *independent* assumption.

Unstructured working correlation specification in GEE modeling assumes different correlations between any two SWA scores for every subject. No constraints are placed on the correlations which are then estimated from the data. Table 4.9 provides the analyses for the parameter estimates for GEE unstructured model. It is clear that, the parameter estimates are the same for all the explanatory variables. A critical observation of the standard errors for empirical and model-based estimators are marginally different and relatively smaller. The variations of these standard error estimates reduce the efficiency of considering the GEE unstructured working correlation assumption as the best fit for the model.

Exchangeable working specification allows for constant correlations between any two measurements within a subject for all the time points. As such, only one parameter needs to be estimated. The working correlation specification therefore allows for constant correlations between any two (2) measurements within a subject i for all time points across the eight (8) semesters.Agaian, the standard errors for the robust and *naïve* cases are marginally different. We can observe from table 4.9 that the results are not different from the previous analyses for independent and unstructured GEE models. This indicates that the true correlation structure for the GEE is not correctly modelled using the *exchangeable* assumption.

Table 4.9 again highlights the model of the academic performance of students using the GEE **Autoregression** (AR-1) working correlation model. The AR-1 measures the correlation within two semesters by their separated time and hence the correlation coefficients diminish as time increases. Similar to the exchangeable model, it requires only one estimated parameter. For the application of GEE models, one assumes that there are fixed number of time-points n that subjects are measured at. The parameter estimates for all the variables are the same for both the empirical and model-based. Also, the standard error estimates for the robust and the model-based are approximately the same. The parameter intercept is generally significant at $\alpha = 0.05$ with the highest standard error in both the empirical and the model-based. The parameter age was significant in the empirical model recording a standard error of 0.276; however, it was insignificant in the model-based model with a standard error of 0.349. Gender, however was not significant in both the empiraical and model-based with a standard error 1.794 of 1.838 and respectively. Entry aggregate was significant in both the robust and the *naïve* parameter estimation. In the parameter estimate for Geographical Location, the difference between the Northern and the Southern Belt (7.136) was significant with standard error of 2.455 and 3.016 in the empirical and model-based parameter estimates. The other contrasts for the Middle and Northern Belts were statistically insignificant and and have their standard errors different in the empirical and model-based parameter estimates. The contrast for the GES categories (A, B and C) were also insignificant and also have the estimation of the standard errors different in the repective empirical and model based.The estimated standard errors for both the robust and *naïve* estimators of the model for all the parameter estimates are marginally equal. In this regard, comparing the analyses of the AR(1) model with the other working correlation assumptions discussed earlier, we chose the AR(1) GEE model as the best fit for our analyses.

4.2.1 AR(1) Model With Linear Time Interactions

Table A.1 in appendix A summ~~arizes the~~ computations for parameter estimates for our GEE working correlation assumptions with respect to specific linear time point interactions across the eight semesters. Again, the parameter estimates for both the empirical and model based were noted to be approximately the same with a significant intercept of 85.566. However there are variations in the standard errors for the robust and *naïve* parameter estimates. Some of the linear trend parameter estimates are negative (decelerating trend). This may imply that, the time interaction effect of the AR(1) model may not be necessarily contributing to the SWA scores of students in the Department. Performance in SWA scores diminishes across time in the linear trend.

The *linear* trend with regards to the main effect for the parameter *age* was not significant in the empirical model as well as the model based estimation. It's parameter estimate was negative (−0.564) indicating a decrease in the average SWA scores over the semesters. The interaction of gender effect is not significant at $\alpha = 0.05$ for both the empirical and model based estimation. The interaction of Geographical Location was not significant in the empirical model; however it was significant in the model based estimation. Entry aggregate was significant with a standard error of 0.456 and 0.383 at $\alpha = 0.05$ level of significance in both the Empirical and Model based estimation respectively.The interaction of GES SHS categories was not significant in both the main effect empirical and model based parameter estimation at $\alpha = 0.05$.

The standard errors for both the empirical and model-based in the AR(1) GEE with Linear Time Interactions model are approximately the same reassuring our preference of considering its working correlation assumptions to the others.

In the Linear Trend time interaction with regards to the GES SHS categories are not significant at $\alpha = 0.05$ level of significance in both the Empirical and Model based estimation respectively. However, in the case of the Geographical location, *M*, *N* and *S* parameter estimates were significant in the empirical based model with standard errors of 0.551, 0.614 and 0.489 respectively but remained insignificant in its corresponding model based parameter estimates.

Table 4.10: Working Correlation Matrix for AR(1) Model With Linear Time

Working Correlation Matrix								
SWA	**SWA1**	**SWA2**	**SWA3**	**SWA4**	**SWA5**	**SWA6**	**SWA7**	**SWA8**
SWA1	1.000	0.798	0.637	0.509	0.406	0.324	0.259	0.206
SWA2	0.798	1.000	0.798	0.637	0.509	0.406	0.324	0.259
SWA3	0.637	0.798	1.000	0.798	0.637	0.509	0.406	0.324
SWA4	0.509	0.637	0.798	1.000	0.798	0.637	0.508	0.406
SWA5	0.406	0.509	0.637	0.798	1.000	0.798	0.637	0.509
SWA6	0.324	0.406	0.509	0.637	0.798	1.000	0.798	0.637
SWA7	0.259	0.324	0.406	0.509	0.637	0.798	1.000	0.798
SWA8	0.206	0.259	0.324	0.406	0.509	0.637	0.798	1.000

The interaction effect between rows and columns variables of the first order authoregressive correlation matrix as seen in Table 4.10 weights the link within two semesters by their estrange time. Statistically, students' SWAs are positively correlated. Generally, there is a slight decrease in the pairwise Pearson Correlation Coefficient over the semesters.

Table A.2 in the appendix summarizes the computations for parameter estimates for our GEE working correlation assumptions with respect to specific *quadratic* time point interactions across the eight semesters. Again, the parameter estimates for both the empirical and model based were noted to be approximately the same with a significant intercept of 91.380. However, there are variations in the standard errors for the robust and *naïve* parameter estimates. Again, some of the quadratic trend parameter estimates are negative (decelerating trend). This may imply that, the time interaction effect of the AR(1) model may not be necessarily contributing to the SWA scores of students in the Department. Performance in SWA scores diminishes across time in the quadratic trend.

The main effect of the parameter *age* in the *quadratic* trend was significant in the empirical model but was not significant in the model based estimation. It's parameter estimate was negative (-0.527) indicating a decrease in the average SWA scores as time increases. The interaction of gender effect is fairly significant at $\alpha = 0.05$ for both the empirical and model based estimation with a *p-value* of 0.060 and 0.053 in the empirical and model based respectively.The interaction of Geographical Location was not significant in both the empirical and model based parameter estimation. Entry aggregate was significant with a standard error of 0.456 and 0.370 at $\alpha = 0.05$ level of significance in both the Empirical and Model based estimation respectively.

With regards to the linear time interaction in the quadratic model Geographical location and GES SHS categories are not significant at $\alpha = 0.05$ level of significance in both the Empirical and Model based estimation respectively. Again, Geographical Location (M, N, and S) are not significant in both the empirical and the model based estimates.

However, in the quadratic trend, the GES SHS categorization A and B was significant in the empirical based model with standard deviations of 0.110 and 0.132 respectively. It however remained insignificant with regards to the parameter estimates in the GES SHS categorization (C) in the model based parameter estimates. Since the linear trend of the GES categorization estimates are negative and its corresponding quadratic estimates are positive, a decelerating negative trend is indicated. GES categorization on SHS diminishes across time in a curvilinear manner.

Table 4.11: Working Correlation Matrix for AR(1) Model With Quadratic Time

Working Correlation Matrix								
SWA	**SWA1**	**SWA2**	**SWA3**	**SWA4**	**SWA5**	**SWA6**	**SWA7**	**SWA8**
SWA1	1.000	0.799	0.640	0.512	0.409	0.328	0.262	0.210
SWA2	0.799	1.000	0.800	0.640	0.512	0.409	0.328	0.262
SWA3	0.640	0.800	1.000	0.800	0.640	0.512	0.409	0.328
SWA4	0.512	0.640	0.800	1.000	0.800	0.640	0.512	0.409
SWA5	0.409	0.512	0.640	0.800	1.000	0.800	0.640	0.512
SWA6	0.328	0.409	0.512	0.640	0.800	1.000	0.800	0.640
SWA7	0.262	0.328	0.409	0.512	0.640	0.800	1.000	0.800
SWA8	0.210	0.262	0.328	0.409	0.512	0.640	0.800	1.000

Again, statistically, students' SWAs are positively correlated. Generally, there is a slight decrease in the pairwise Pearson Correlation Coefficient over the semesters. It is also observed that all the Pearson Correlation Coefficients are strong (> 0.5). However, the relationship is strongest between immediate semesters.

4.2.2 Hypothesis Testing

In consideration to the outputs revealed from SAS, we test the appropriateness or otherwise of which of the models (AR(1) Linear Trend Model or AR(1) Quadratic Trend Model) is the best fit statistical model.

The appropriate hypothesis is given as

$H_0 = AR(1)$ Linear Trend Model fits data better

$H_1 : AR(1)$ Quadratic Trend Model fit data better

With a critical value of approximately 14 and a chi-square $\chi^2{}_{(0.05,7)} = 20.28$, we fail to reject the *Null Hypothesis* and therefore conclude that the AR(1) model with the Linear Trend is a preferred statistical model than the AR(1) Quadratic Trend Model.

4.2.3 Random Intercept Model

The model below is exactly what we would use in deciding as to whether or not to select a random effect model for the data. The model contains only one parameter which is the random intercept effect. This partition the total variation in the data into two: within individual and between individual components

$$SWA_{ij} = \upsilon_{oi} + \varepsilon_{ij} \tag{4.1}$$

where

$$\upsilon_{0i} \sim N(\sigma^2_{r0})$$

and

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

A useful tool used in deciding whether a random effect model would be an appropriate choice for the data is the *Intraclass Correlation Coefficient* (ICC) and it is represented mathematically below

$$ICC = \frac{\widehat{\sigma^2}_{\upsilon_o}}{\widehat{\sigma^2}_{\upsilon_o} + \widehat{\sigma^2}} \tag{4.2}$$

where $\widehat{\sigma^2}$ is the residual variance

From Table 4.12, $\widehat{\sigma^2}_{\upsilon_o} = 39.876$ & $\widehat{\sigma^2} = 19.939$

Therefore, it implies that our $ICC = \frac{39.876}{39.876+19.939} = 0.666$, indicating that approximately 67% of the variation in the data is explained by allowing the intercept to vary across the

Table 4.12: Covariance Parameter Estimates

COV PARM	SUBJECT	ESTIMATE
UN(1,1)	Student's Identity	39.876
RESIDUAL		19.939

individual students. The statistical significant value for the within individual variation suggests the data structure is best captured by using *Random Effect Model.* With a covarinace value of 39.876 and a residual of 19.939, it is clear that, there is a lot of variations in the academic performance of students at semester one (1).

With regards to Table 4.13, the *random intercept model* is considered. It is a linear mixed model where only subject specific effect is the intercept. Statistically, not all the parameter estimates are significant at $\alpha = 0.05$ level of significance. It was observed that geographical location was not significant. However, entry aggregate, entry age and gender were all significant. On the average, there is no significance difference in the SWA between the GES type of schools over a linear trend as well as the middle and southern belts with regards to the linear trends in the geographical location. Nevertheless, the geographical location with regards to the Northern Belt was significant in the linear trend with a standard deviation of approximately 0.742.

In the Type (III) Tests, entry age, entry aggregate and linear trend in the geographical location of the students are significant.

Table 4.13: Random Effect or Mixed Model Random Intercept

Solution for Fixed Effects								
EFFECT	**GENDER**	**REGION**	**GES.CAT.**	**ESTIMATE**	**STAND. ERR**	**DF**	**t VALUE**	**PR $\geq \lvert t \rvert$**
INTERCEPT				85.894	9.894	65	8.68	$\geq$ 0.0001
AGE				-0.566	0.283	509	-2.00	0.046
GENDER	F			86.500	9.887	509	8.75	$\geq$ 0.0001
GENDER	M			0	.	.	.	.
REGION		M		1.917	1.529	509	1.25	0.210
REGION		N		3.628	2.958	509	1.23	0.221
REGION		S		0	.	.	.	.
ENTRY AGG.				-1.310	0.477	509	-2.74	0.006
GES.CAT			A	-2.067	1.563	509	-1.32	0.187
GES.CAT			B	1.186	1.882	509	0.63	0.529
GES.CAT			C	-1.849	2.555	509	-0.72	0.470
GES.CAT			P	0	.	.	.	.
TIME*REGION		*M		1.099	0.661	509	1.66	0.097
TIME*REGION		*N		1.798	0.712	509	2.53	0.012
TIME*REGION		*S		0.832	0.591	509	1.41	0.160
TIME*GES.CAT			*A	0.147	0.659	509	0.22	0.824
TIME*GES.CAT			*B	-0.194	0.661	509	-0.29	0.769
TIME*GES.CAT			*C	0.541	0.667	509	0.81	0.417
TIME*GES.CAT			*P	0.000	.	.	.	.

4.2.4 Random Intercept & Slope Model

Again, consider the model below:

$$SWA_{ij} = \upsilon_{oi} + \upsilon_{1i}t_{ij} + \varepsilon_{ij} \tag{4.3}$$

where

$$\upsilon_{0i} \sim N(\sigma^2_{r0})$$

$$\upsilon_{1i} \sim N(\sigma^2_{r1}),$$

and

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

The ICC is computed using variance estimates for the random intercept and slope model as in the equation above as well as their covariances.

Table 4.14: Covariance Parameter Estimates

COV PARM	SUBJECT	ESTIMATE
UN(1,1)	Student's Identity	29.594
UN(2,1)	Student's Identity	-0.450
UN(2,2)	Student's Identity	0.778
RESIDUAL		15.613

With regards to Table 4.14, $ICC = \frac{29.594-0.450+0.778}{29.594-0.450+0.778+15.613} = 0.657$, indicating that approximately 65.7% of the variation in the data is accounted for by allowing the intercept and slope to vary across individual students in Maths four. The random intercept has a relatively large estimate with respect to the other variance components. This supports the fact that there is high difference between student's variability based on SWA's at semester 1. The negative estimate of the covariance implies that students who start with a high SWA at semester 1, have a more tendency to exhibit reduction of SWA over the semesters. Variation within the slope and the semester was however not significant. Therefore variations in students academic performance as the semester progresses is very minimal. Hence can be ignored in the final model.

In Table B.1 in the Appendix, the *random intercept and slope model* together with some regressors and individual slopes are considered. It is a linear mixed model where only subject specific effect is the intercept. Again, not all the parameter estimates are significant at $\alpha = 0.05$ level of significance. It was observed that *geographical location* in the main model as well as the linear trend in the *GES categorization* type of SHS school were not significant. However, the parameter *intercept, entry age, gender* and *entry aggregate* were all significant. On the average, there is no significance difference in the SWA between the *geographical location*(M,S) over a linear trend. Nevertheless, the *geographical location* with regards to the *northern belt* was significant in the linear trend with a standard deviation of approximately 0.749. Again, in the Type (III) Tests, *entry age, entry aggregate* and linear trend in the *geographical location* of the students are significant.

Chapter 5

CONCLUSION AND RECOMMENDATIONS

5.1 Discussion

The study attempts to find the factor(s) that can have impact or affect academic performance. A measure of academic performance used was SWA scores obtained over the eight (8). The main factors considered are entry age, gender,entry aggregate, GES graded level of SHS attended and geographical location. Whether these socio-dermographic factors affect students Academic Performance is the researcher's priority. The statistical model used here were the GEE model and the Random Effect Model.

As part of the exploratory data analysis, it was observed that, out of a total of 90 students, 15 students were females representing 16.85% and 74 students were males representing 83.15%. The Geographical Location of Maths Four (4) students as in was grouped into three zonal belts specifying their respective regions of origin. These include the Northern Belt (includes Northern, Upper East, Upper West), Middle Belt (Ashanti, Brong Ahafo

Regions, Eastern, Volta) and Southern/Coastal Belt (includes Greater Accra, Central and Western regions) have been specified. Fifty (50) of the students representing 58.14% represented the Middle Belt. Also, ten (10) of the students representing 11.63% represented the Northern Belt. Similarly, twenty-six (26) of the students representing 30.23% represented the Southern Belt. The Geographical Location of four (4) students was not captured in the data and therefore was treated as missing.Again,the students were grouped according to the GES special SHS categorization with respect to the type of SHS attended; whether Grade A, B, C, D or P schools were specified. It could be observed that out of the total of 90 students, 47 students (representing 55.29%) attended a Grade A schools, 25 students (representing 29.41%) attended Grade B school, 9 students (representing 10.59%) attended a Grade C school, 2 students (representing 2.35%) attended both a Grade D school and a Grade P school respectively, whilst 5 students information failed to be captured and therefore was recorded as missing.

With regards to the GEE estimation, an analysis is carried out by assuming a priori a certain 'working correlation structure for the repeated measurements of the outcome variable. Four (4) correlation assumptions were considered namely: *independent*, *exchangeable*, *autoregressive* and *unstructured* correlation structure.The study reaffirms the consisent estimate of GEE with its corresponding working correlation assumption matrix. The results of sitting the model parameter estimates are approximately identical, however, the standard errors for the various GEE model varies within and across the parameters. This again confirms the closeness of the measures for comparison of students mean score in the Department. This closeness in the standard errors in both the empirical and the model based parameter estimates in the GEE correlation assumption for AR(1) confirms its suitability for the true regression model.

The *linear* trend with regards to the main effect for the parameter *age* was not significant in the empirical model as well as the model based estimation. It's parameter estimate was negative (-0.564) indicating a decrease in the average SWA scores as time increases. The interaction of *gender* effect is not significant at $\alpha = 0.05$ for both the empirical and model based estimation. The interaction of *geographical location* was not significant in the empirical model; however it was significant in the model based parameter estimation. *Entry aggregate* was significant with a standard error of 0.456 and 0.383 at $\alpha = 0.05$ level of significance in both the empirical and model based estimation respectively.The interaction of *GES SHS categories* was not significant in both the main effect empirical and model based parameter estimation at $\alpha = 0.05$.

The standard errors for both the empirical and model based in the AR(1) GEE with Linear Time Interactions model are approximately the same reassuring our preference of considering its working correlation assumptions to the others.

In the Linear Trend time interaction with regards to the *GES SHS categories* are not significant at $\alpha = 0.05$ level of significance in both the empirical and model based estimation respectively. However, in the case of the *geographical location*, *M*, *N* and *S* parameter estimates were significant in the empirical based model with standard errors of 0.551, 0.614 and 0.489 respectively but remained insignificant in its corresponding model based parameter estimates.

In longitudinal studies the observations within one subject over time are correlated. The observations over time are nested within the subject. The basic idea behind the use of multilevel techniques in longitudinal studies is that the regression coefficients are allowed to differ between subjects. Therefore the term random coefficient analysis is preferred to the term multilevel analysis. The simplest form of random effects analysis is an analysis with only a random intercept. It is also possible that the intercept is not random, but that the development of a certain variable over time is allowed to vary among subjects or, in other words, the slope with time is considered to be random. For longitudinal studies, random effects models enable the analyst to not only describe the trend over time while taking into account of the correlation that exist between massive measurements, but also to describe the variation in the baseline measurement and in the rate of change over time.

Since the interaction or variation around the slope was highly insignificant, the *random intercept model* was the better alternative ahead of the *random intercept and slope model.* The *random intercept model* is a linear mixed model where only subject specific effect is the intercept. Statistically, not all the parameter estimates are significant at $\alpha = 0.05$ level of significance. It was observed that, the difference in *geographical location* was not significant in the main effect model. However, *entry aggregate*, *entry age* and *gender* were all significant. On the average, there is no significant difference in the SWA between the *GES SHS categories* type of schools over a linear trend as well as the *middle* and *southern belts* with regards to the linear trends in the *geographical location.* Nevertheless, the *geographical location* with regards to the *northern Belt* was significant in the linear trend with a standard deviation of approximately 0.742.

Again, with regards to the *Type (III)* Tests, *entry age*, *gender*, *entry aggregate* and linear trend in the *geographical location* of the students are significant.

5.2 Conclusion

The parameter estimate of the *entry aggregate* in the Linear Trend of the AR(1) GEE Model was *negative*; implying that, a lower entry aggregate will have an approximate average of 33% increase in SWA.

In addition, it is observed that, students from the *Northern* Belt on an average with regards to SWA score increase approximately by 1.7% per semester whereas students from the *Middle* and *Southern* Belt increase by 1.2% and 0.9% per semester respectively.

Again, on the average, students with lower entry ages have on a higher SWA score.

The *GEE AR(1) Linear Trend Model* and the Fixed Effect results from the *Random Intercept Model* showed the samilar significant factors ie. age , entry aggregate and geographical location.

5.3 Recommendation

After careful analysis of the study

- With regards to subsequent research into this area, other factors on the part of students such as fee-paying or regular, foreign or local students, etc should be taken into consideration. Questionaires can also be designed to enquire about a student's time spent at the library in studying, attending lectures etc.

References

Adams, A. (1996). Even basic needs of young are not met. *Falk School Library.*

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

Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis on Count Data.* New York: Cambridge University Press.

Caprara, G., Vecchione, M., Alessandri, G., Gerbino, M., and Barbaranelli, C. (2011). The contribution of personality traits and self-efficacy beliefs to academic achievement: A longitudinal study. *British Journal of Educational Psychology*, 81:78–96.

Chamorro-Premuzic, T. and Furnham, A. (2003). Personality predicts academic performance: Evidence from two longitudinal university samples. *Journal of Research in Personality*, 37:319–338.

Chandler, R. and Wheater, H. (2002). Analysis of rainfall variability using generalized linear models: A case study from the west of ireland. *Water Resources Research*, 38(10).

Diggle, P. and Kenward, M. G. (1994). Informative drop-outin longitudinal data analysis. *Appl. Statist*, 43(1):49–93.

Felder, R., Felder, G., Mauney, M., Hamrin Jr., C., and Deitz, E. (1995). A longitudinal study of engineering student performance and retention. iii. gender differences in student performance and attitudes. *Journal of Engineering Education*, 84(2):151–163.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). *Applied Longitudinal Analysis.* Wiley New York.

Gibson, B. and Cassar, G. (2005). Longitudinal analysis between relationships planning and performance in smallfirms. *Small Business Economics*, 25:207–222.

Goldstein, H. (1991). *Multilevel statistical models.* London: Edward Arnold.

Guay, F., Larose, S., and Boivin, M. (2004). Academic self-concept and educational attainment level: A ten-year longitudinal study. *Self and Identity.*

Horwood, L. and Fergusson, D. (1999). A longitudinal study of maternal labour force participation and child academic achievement. *J. Child Psychol. Psychiat*, 40(7):1013–1024.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression.* Wiley New, 2nd edition.

Hoy, W., Tarter, C., and Hoy, ~~A. (2006)~~. Academic optimism of schools: A force for student achievement. *American Educational Research Journal*, 43(3):425–446. Published by: American Educational Research Association.

Jhermanowicz, J. C. (2003). Scientists and satisfaction. *Social Studies of Science*, 33(1):45–73. Sage Publications, Ltd.

Jimerson, S., Ferguson, P., Whipple, A., Anderson, G., Dalton, M., et al. (2002). Exploring the association between grade retention and dropout: A longitudinal study examining socio-emotional, behavioral, and achievement characteristics of retained students. *The California School Psychologist*, 7:51–62.

Kiamanesh and Kheirieh (2001). Trends in mathematics educational inputs and outputs in iran: Findings from the third international mathematics and science study and its repeat. *Tehran: Institute for Educational Research Publication.*

Kohli, T. K. (1975). Characteristic behavioural and environmental correlates of academic achievement of over and under achievers at different levels of intelligence. Punjab University, unpublished Ph.D. Thesis. Page. 48.

Laird, N. and Ware, J. (1986). Random effects models for longitudinal data. *Biometrics*, 38:963–974.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–122.

Liang, K. Y. and Zeger, S. L. (1993). Regression analysis for correlated data. *Annual Review of Public Health*, 14:43–68.

Lipatz, S. R., Fitzmaurice, G. M., Orav, E. J., Laird, N. M., et al. (1994a). Performance of generalized estimating equations in practical situations. *Biometrics*, 50:270–278.

Longford, N. T. (1993). *Random coefficient models.* Oxford University Press.

Loyalka, P., Song, S., Wei, J., Rozelle, S., et al. (2010). Information, college decisions and financial aid: Evidence from a cluster-randomized control trial in china. *China Institute for Educational FInance Research.*

Marsh, H. and Yeung, A. (1998). Longitudinal structural equation models of academic self-concept and achievement: Gender differences in the development of math and english constructs. *American Educational Research Journal*, 35(4):705–738.

McCombs, B. L. and Marzano, R. J. (1990). Putting the self in self-regulated learning: The self as agent in integrating will and skill. *Educational Psychologist*, 25:51–69.

McCullagh, P. and Nelders, J. A. (1989). *Generalized Linear Models.* Chapman and Hall, 2nd edition.

McLemore, C. (2010). Resiliency and academic performance. *ScholarCentre.*

M.Spelling, editor (2005). *Confidence: Helping your child through early adolescence.* ED Pubs.

Nelders, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, pages 370–384. Series A.

Pannenberg, M. and Spiess, M. (2007). Gee estimation of a two-equation panel data model: An analysis of wage dynamics and the incidence of profit-sharing in west germany.

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.

Park, T. (1993). A comparison of generalized estimating equation approach with the maximum likelihood approach for repeated measurements. *Statistics in Medicine*, 12:1723–1732.

Preisser, J., Young, M., Zaccaro, D., Wolfson, M., et al. (2003). An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. *Stat Med. 2003 Apr 30;*, 22(8):1235–1254.

Rogosa, D. and Saner, H. (1995). Longitudinal data analysis examples with random coefficient models. *Journal of Educational and Behavioral Statistics*, 20(2):149–170.

Royall, R. M. (1986). Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review*, 54:221–226.

Scalesa, P., Bensona, A., Roehlkepartaina, E., Arturo, S. J., van Dulmen, M., et al. (2006). The role of developmental assets in predicting academic achievement: A longitudinal study. *Journal of Adolescence*, pages 691–708.

Schaible, G., Kim, S., and Lambert, D. (2007). *Structural Conservation Practices in U.S. Corn Production: Evidence on Environmental Stewardship by Program Participants and Non-Participants*. Economic Research Service, USDA.

Schneider, B. and Lee, Y. (1990). A model for academic success: The school and home environment of east asian students. *Anthropology & Education Quarterly*, 21(4):358–377. Published by: Blackwell Publishing on behalf of the American Anthropological Association.

Stewart, S. M., Lam, T. H., Betson, C., Wong, C., Wong, A. M., et al. (1999). A prospective analysis of stress and academic performance in the first two years of medical school. *Med Edu*, 33:243–50.

Teachman, J. (1997). Gender of siblings, cognitive achievement, and academic performance: Familial and nonfamilial influences on children. *Journal of Marriage and Family*, 59(2):363–374. Published by: National Council on Family Relations.

Twisk, J. W. R., Staal, B. J., Brinkman, M. N., Kemper, H. C. G., van Machelen, W., et al. (1998a). Tracking of lung function parameters and the longitudinal relationship with lifestyle. *European Respiratory Journal*, 12:627–634.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61(3):439–447.

Zajacova, A., Lynch, S., and Espenshadet, T. (2005). Self-efficacy, stress, academic success in college. *Research in Higher Education*, 46(6).

Zeger, S. L. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcome. *Biometrika*, 42:121–130.

Zeger, S. L. and Liang, K. Y. (1992). An overview of methods for the analysis of longitudinal. *Statistics in Medicine*, 11:1825–1839.

Zeger, S. L., Liang, K. Y., Albert, P. S., et al. (1988). Models for longitudinal data: A generalized estimating equations appraoch. *Biometrics*, 44:1049–1060.

Appendix A

Results of Linear and Quadratic Trends

Table A.1: AR(1) Model with Linear Time & Standard Errors

		ESTIMATES	MODEL BASED	PR $\geq \lvert Z \rvert$
Parameters	Codes		Std Err	
INTERCEPT		85.566	10.849	$\leq$ 0.0001
AGE		-0.564	0.332	0.034
GENDER	F	3.271	1.751	0.067
GENDER	M	0.000	0.000	.
REGION	M	1.297	2.230	0.445
REGION	N	3.500	3.918	0.255
REGION	S	0.000	0.000	.
ENTRY.AGG		-1.320	0.383	0.004
GES.CAT	A	0.257	6.226	0.886
GES.CAT	B	3.211	6.382	0.118
GES.CAT	C	0.658	6.989	0.806
GES.CAT	P	0.000	0.000	.
TIME*REGION	*M	1.227	1.030	0.026
TIME*REGION	*N	1.701	0.860	0.006
TIME*REGION	*S	0.929	1.134	0.058
TIME*GES.CAT	*A	0.052	0.964	0.924
TIME*GES.CAT	*B	-0.250	1.014	0.651
TIME*GES.CAT	*C	0.284	1.041	0.615
TIME*GES.CAT	*P	0.000	0.000	.

(Std Err) Shows Standard Error

Table A.2: AR(1) Model with Quadratic Time & Standard Errors

Parameters	Codes	ESTIMATES	MODEL BASED Std Err	PR ≥ $\|Z\|$
INTERCEPT		91.380	8.678	≤ 0.0001
AGE		-0.543	0.313	0.083
GENDER	F	3.240	1.677	0.053
GENDER	M	0.000	0.000	.
REGION	M	0.444	2.431	0.855
REGION	N	0.477	4.305	0.912
REGION	S	0.000	0.000	.
ENTRY.AGG		-1.320	0.370	0.0004
TIME*REGION	*M	0.339	2.266	0.881
TIME*REGION	*N	2.182	2.675	0.415
TIME*REGION	*S	-0.947	2.163	0.661
TIME*GES.CAT	*A	-2.689	2.191	0.220
TIME*GES.CAT	*B	-1.997	2.247	0.374
TIME*GES.CAT	*C	-1.642	2.457	0.504
TIME*GES.CAT	*P	0.000	0.000	.
TIMESQ*REGION	**M	0.030	0.280	0.915
TIMESQ*REGION	**N	-0.118	0.317	0.711
TIMESQ*REGION	**S	0.148	0.264	0.574
TIMESQ*GES.CAT	**A	0.367	0.273	0.180
TIMESQ*GES.CAT	**B	0.281	0.281	0.316
TIMESQ*GES.CAT	**C	0.264	0.306	0.388
TIMESQ*GES.CAT	**P	0.000	0.000	.

(Std Err) Shows Standard Error

Appendix B

Random Intercept & Slope Model

Table B.1: Random Effect or Mixed Model Random Intercept & Slope

Solution for Fixed Effects								
EFFECT	GENDER	REGION	GES.CAT.	ESTIMATE	STAND. ERR	DF	t VALUE	PR $\geq \lvert t \rvert$
INTERCEPT				90.745	8.561	65	10.60	$\geq$ 0.0001
AGE				-0.670	0.313	441	-2.14	0.033
GENDER	F			1.271	1.864	441	0.68	0.496
GENDER	M			0	.	.	.	.
REGION		M		2.215	1.526	441	1.45	0.148
REGION		N		3.940	2.977	441	1.32	0.186
REGION		S		0	.	.	.	.
ENTRY AGG.				-1.498	0.366	441	-4.09	$\geq$ 0.0001
GES.CAT			A	-1.730	1.496	441	-1.16	0.248
GES.CAT			B	1.346	1.805	441	0.75	0.456
GES.CAT			C	-1.734	2.545	441	-0.68	0.496
GES.CAT			P	0	.	.	.	.
TIME*REGION		*M		1.119	0.660	441	1.70	0.091
TIME*REGION		*N		1.820	0.711	441	2.56	0.011
TIME*REGION		*S		0.832	0.591	441	1.41	0.160
TIME*GES.CAT			*A	0.123	0.657	441	0.19	0.852
TIME*GES.CAT			*B	-0.209	0.660	441	-0.32	0.752
TIME*GES.CAT			*C	0.524	0.666	441	0.79	0.432
TIME*GES.CAT			*P	0	.	.	.	.