



UNIVERSITY OF NAIROBI
SCHOOL OF MATHEMATICS

Modeling Income Distribution of Households in Mwea: A Hierarchical linear Model approach

This research project is submitted to the School of Mathematics of the University of Nairobi in partial fulfillment of the requirement for the degree of Masters of Science in Social Statistics.

By

KABUU NJOROGE JOHN

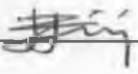
July 2012

DECLARATION

This project as presented in this report is my original work and has not been presented for any other university award

Student:

Kabuu Njoroge John Reg. No.: I56/P/7918/2003

Signature:  _____

Date: 30 / 7 /2012

This project has been submitted as a partial fulfillment of the requirements for Masters of Science in Social Statistics at the University of Nairobi with our/my approval as the university supervisor(s).

Mr. John Ndiritu

Signature:  _____

Date: 30 / 07 /2012

ACKNOWLEDGEMENT

My sincere regards to my lovely parents Mr. & Mrs. Peter Kabuu for their support that made my life better in countless ways. Special thanks to the Directors of KIU Constructions Mr & Mrs Kariuki Theuri for moral support, mentorship and financial assistance, the Director of O'cra Company limited Eng. Bernard Maina for provision of the Data used in this project and finally I do honour my supervisor Mr. John Ndiritu for his tireless contribution and guidance in the course of my project working.

ABSTRACT

Mwea Irrigation Scheme having been formed as early as 1956, the project studies the current distribution of Gross family income in the sections of Mwea Irrigation Scheme accounting for individual's characteristics of the persons. Different kinds of HLM models were used to get a deeper understanding of the variations in Gross family income both within and across the sections that constitute Mwea Irrigation Scheme. The significant findings of the project were that; there was non-significant difference in Gross family income between the individuals within and across the sections with respect to the effects of the measured individual's characteristics under study. The study found a significant difference in Gross family income distribution across the sections. Subject to our discussion, the research findings suggests that there are other factors other than the ones analyzed in this study that result into sectional differences in Gross family income and hence a need to perform further research.

Key Words: Hierarchical Linear models.

TABLE OF CONTENTS

DECLARATION	II
ACKNOWLEDGEMENT	III
ABSTRACT	IV
TABLE OF CONTENTS.....	V
LIST OF FIGURES	VII
LIST OF TABLES	VIII
ABBREVIATIONS AND ACRONYMS.....	IX
CHAPTER ONE.....	1
INTRODUCTION.....	1
1.1 BACKGROUND INFORMATION.....	1
1.2 THE ROLE OF NIB.....	1
1.3 PROBLEM STATEMENT	3
1.4 GENERAL OBJECTIVE	3
CHAPTER TWO.....	5
LITERATURE REVIEW.....	5
2.1 INTRODUCTION.....	5
2.2 THE ROLE OF INCOME VARIATIONS	5
2.2 OVERVIEW OF HIERARCHICAL LINEAR MODELING (HLM).....	7
CHAPTER THREE	11
METHODS	11
3.1 INTRODUCTION	11
3.2 RESEARCH DESIGN.....	11
3.2.1 <i>Stratified sampling</i>	12
3.2.2 <i>Cluster sampling</i>	12
3.3 HIERARCHICAL LINEAR MODELS	13
3.3.1 <i>Introduction</i>	13
3.3.2 <i>Justification in the use of HLM(s)</i>	14
3.3.3 <i>A 2 level HLM model for continuous response variable and model Assumptions</i>	15
3.3.4 <i>Fitting a 2 level hierarchical linear model</i>	16
3.3.4.1 Fitting a null model (empty model)	16
3.3.4.2 Fitting a Random intercept model.....	18
3.3.4.3 Fitting a random slope model.....	19
3.3.4.4 Fitting a Random intercept and random slope model	20
3.3.3.4 The compressed case for the hypothesized models.....	22
3.3.5 <i>Model Assumptions</i>	24
3.3.6 <i>Hypothesis Testing (parameter testing's)</i>	25
3.3.6 <i>Testing the fit of the model</i>	25
3.3.7 <i>Estimation procedure</i>	25

3.4 MEASUREMENT INSTRUMENTS	26
3.5 STATISTICAL SOFTWARE.....	26
3.6 HYPOTHESIZED MODELS IN THE STUDY	27
3.6.1 <i>The null model or unconditional model</i>	27
3.6.2 <i>A model incorporating the demographic factors as fixed effects</i>	28
3.6.3 <i>A model incorporating the institutional factors as fixed effects</i>	29
3.6.4 <i>A model having demographic factors as random effects</i>	30
CHAPTER FOUR.....	31
RESULTS.....	31
4.1 DATA.....	31
4.1.2 <i>Data description and variable specifications</i>	31
4.1.3 <i>Data preparation</i>	32
4.2 ANALYSIS AND RESULTS INTERPRETATIONS	33
4.2.1 <i>Model fitting</i>	34
4.2.1.1 Fitting the null model	35
4.2.1.2 Modeling demographic factors.....	37
4.2.1.3 Modeling institutional factors	49
CHAPTER FIVE.....	53
CONCLUSION AND DISCUSSIONS	53
REFERENCES.....	56

LIST OF FIGURES

Figure 1: A histogram for the log (gross family income)	33
Figure 2: Frequency table showing the percentage of the house holds who accessed credit in all the MIS sections.....	34
Figure 3: Frequency table for the level of education	34
Figure 4: Frequency table for the extension services.....	34
Figure 6: A plot of the estimated income against gender for independent random intercept random slope model	40
Figure 7: A plot of the estimated variance against age	42
Figure 8: A plot of estimated income against age.....	43
Figure 9: A plot of the estimated income against level of education.....	46
Figure 10: A plot of estimated gross family income against extension services	50
Figure 11: A plot of estimated income against access credit	52

LIST OF TABLES

Table 1: Sections of the MIS and total number of beneficiaries per section.	11
Table 2: Results for the null model.	35
Table 3: Sectional effects and their respective standard errors.....	35
Table 4: Results for the null single level model.....	37
Table 5: Multilevel model with gender as the explanatory variable.....	37
Table 6: Results for the independent random intercept random slope model.....	40
Table 7: Multilevel models with age as the explanatory variable.....	41
Table 8: Results for models hypothesis testing.....	42
Table 9: Multilevel models with level of education as the explanatory variable.....	44
Table 10: Multilevel models with marital status as the explanatory variable.....	47
Table 11: Fixed part of multilevel models with demographic factors as the explanatory variable	48
Table 12: Multilevel models with extension services as the explanatory variable.....	49
Table 13: Multilevel models with access to credit as the explanatory variable.....	51

ABBREVIATIONS AND ACRONYMS

AIC	Akaike's Information Criterion
ANOVA	Analysis of Variance
BIC	Bayes Information Criterion
HLM	Hierarchical Linear Models
ICC	Intra Class Correlation
LRT	Likelihood Ratio Test
MIS	Mwea Irrigation Scheme
MLR	Multilple Linear Regression
NIB	National Irrigation Board
OLS	Ordinary Least Square Regression
PSU	Primary Sampling Unit
SRS	Simple Random Sampling
SSU	Secondary Sampling Unit
WUA	Water User's Association

CHAPTER ONE:

INTRODUCTION

1.1 Background Information

Mwea Irrigation Scheme (MIS) was initiated in 1956 in Kirinyaga District of Central Province as a paddy production scheme to provide means of food and livelihood to resettled 'formerly landless' people. From an initial 2,000 acres, the scheme has undergone expansion to the current 30,000 acres. A total of 16,000 acres has been developed for paddy production and the rest of the scheme is used for settlement, public utilities, subsistence and horticultural crops farming. Between 1999 and 2003 there was proliferation of unplanned and un designed expansions of irrigated areas along the irrigation waters conveyance, distributions, and drainage channels, and this led to depilated state of the scheme.

The scheme is subdivided into sections namely; Tabere with 1100 beneficiaries, Karama with 749 beneficiaries ,Mwea with 928 beneficiaries ,Thiba with 868 beneficiaries ,Wamumu with 843 beneficiaries ,Mutithi with 1761 beneficiaries, Marura with 280 beneficiaries , Nderwa north with 727 beneficiaries. The subdivision aims to achieve efficient management and service delivery. The scheme is served by two major rivers; rivers Nyamidi and Thiba, and water for irrigation abstracted from rivers by gravity by use of fixed intake weirs, conveyed and distributed in the scheme. The scheme is run by National Irrigation Board (NIB) and the farmer's organizations mainly Water user's Association (WUA).

1.2 The role of NIB

NIB is the largest player in the large scale irrigation subsector in Kenya and it's expected to play a leading role in the management, coordination and implementation activities. NIB assumes the role of leadership in public relations in the irrigation and drainage subsector, developing appropriate mechanisms and guidance for implementing agencies including monitoring and evaluation, collaborate with national and international agencies in irrigation development, prepare proposals and estimates to be submitted to government and other potential donors and collaborators, resource mobilizations for rehabilitation and development of all irrigation and

drainage activities in the irrigation schemes. At the scheme level, scheme managements are expected to provide proactive leadership in the implementation plan.

The work of NIB in the scheme is to improve the standard of living of the rural population by ensuring food security through production of quality food products and cash crops that would generate reasonable incomes for the Kenyan population, create wealth and employment by supporting farmers to expand their farming activities and output. Jobs are created at the farm level as well in processing and marketing of agricultural products.

The National Irrigation Board corporate plan objectives include issues of normalization of operations in the existing irrigation schemes through rehabilitation and irrigation, drainage infrastructure, sensitization and capacity building of farmers so as to increase the crop production and hence income level of the farmers.

The current rice production from MIS is estimated to be 43,000 Tonnes and the government has plans to rehabilitate MIS in the next three years, and estimates there will be a double increase of crop production output. According to observation assessment of this area, a substantial percentage of the residents of MIS depend on agricultural cultivation of rice and horticultural farming as their primary occupation and a small percentage of them have engaged in business activities as their second occupational option.

Agricultural activities which constitute crop farming and agricultural labor on the farms form the main source of income for these tenants. Agricultural labor on these farms, form a means of livelihood for the tenants and a source of family income. Other sources of income, where we term it as non farm income, comprises: salary from other occupation (as an employee) such as government employee, pensions; wages for casual work; earnings from sale of fuel wood and forest products; earnings from sale of fodder products; earnings from livestock and poultry products; earnings from tree crops (own trees and other's trees); earnings from home garden; earnings from subsidiary business after deducting the cost for the business or service; receipt of gifts and remittances from relatives and others; income from lending charges for farm

machinery, equipment and work animals after deducting cost for the service; income from leased land; interest earned from money loaned to other persons and bank deposits.

1.3 Problem statement

The scheme having started as early as 1956 and the government involvement in its development and expansion in terms of providing infrastructural facilities such as provision of market accessibility, capacity building and the objective core values of NIB, the prime question to resolve is how the gross family income of households in these sections are distributed and what factors influence Gross family income distributions in these sections.

1.4 General objective

In the current study, we intend to study and determine whether demographic factors such as Age of the house hold head, Marital status of the respondent usually the household head, level of education of the household head, gender of the household head in addition to institutional factors like access to credit and availability of agricultural extension services poses any significant effect in the distribution of family income in Mwea Irrigation Scheme (MIS).

1.5 Specific objectives

The study seeks to investigate:

- i. Whether there is difference in the distribution of Gross family income across the MIS sections.
- ii. Whether demographic factors have significant effect in the variability of Gross family income across the MIS sections.
- iii. Whether institutional factors have significant effect in the variability of Gross family income across the MIS sections.

1.6 Justification

An understanding of how a society's distribution of income and the related variances in there distributions is of prime importance in setting up development programs or enforcing the implementation of the existing programs so as the society can benefit at large. For the case of an agricultural setting like MIS, where the primary occupation is Rice production, an understanding of income distribution in this area will assist the government in formulating policies that can not

only boost the well being of these farmers in increasing production, but can also, increase food security for the nation.

1.7 Limitations

Time was a limiting factor in this study. There are other areas of the study we intend to explore and especially the area related to disease and crop husbandry, human social economic factors in relationship to income levels of the families.

In the proceeding chapter, we will look at studies done in relation to income inequality and distribution in relation to factors that affect income distributions. An introduction of the method used in this research work and an in depth understanding of how the models are formulated will be discussed in chapter 3. Chapter4 provides the results of all the hypothesized models formulated in chapter 3 and there is a substantial amount of information in relation to the explanatory variables used in the study. Chapter 5 elaborates an evaluative discussion and conclusion of the study.

CHAPTER TWO:

LITERATURE REVIEW

2.1 Introduction

In the following chapter, we intend to review the works done in relation to income studies and later have an overview of the mathematical technique to be used in the analysis of this study.

2.2 The Role of Income variations

Income variations or inequality has been an intense subject of study for most of researchers in the world. Some of these studies have evaluated such geographical units as continents to countries up to the least geographical unit of community level. Income inequality within the level of household and even states are affected by a variety of factors, and according to (Terry *et al*, 2001) notes that a measure by a Gini and Paglin Concentration ratios is related to personal characteristics such as schooling, Occupation, Age, Sex ,Race and Labor market characteristics such as Region, Unemployment and Industrial structure.

The inequality studies generally use the median years of schooling as an explanatory variable, and find that inequality decreases as the level of schooling increases. As (Terry *et al*, 2001) concludes that dispersion increases with skill level as well as inter age dispersion due to steeper income profiles. On the other hand, the percentage of female headed house holds and families with heads aged 55 to 64 years were primarily associated with greater intra age dispersion income.

The government has in the past attempted to solve the problems of poverty and income inequality through development plans and strategies. Such plans and strategies are like the Sessional Paper No. 10 of 1965 on African socialism and its application to planning. It advocated the development of a dual economy through unbalanced investments, with more investment in areas expected to yield the largest net output. Thus areas having abundant natural resources, favorable climate, developed infrastructure and people more receptive to, and active in development were favored as noted by (Gitau, 2005).

Other major measures included the adoption of the “Basic Needs Approach” In the 1970’s and the Sessional Paper No.1 of 1986 on economic management for reviewed growth. The former involved heavy investments in education and Health as a tacit way of dealing with disparities, but suffered from poor implementation, resulting in poor impacts. Starting in the early 1990’s, the country adopted much of the Market liberalization and structural adjustment reforms that swept the world, with less explicit attention paid to poverty than in the past.

According to the most recent National population study 2009, shows that National poverty is in the down ward trend, falling from 56% in 2000 to 46%, where poverty has been defined and measured in Economic welfare terms such as income or consumption (*Kenya Demographic and Health Survey report, 2010*). A study by (Mathenge & Tschirley, 2008) analyses house hold income growth and mobility with emphasis on education contributions and poverty persistence.

The studies conducted on inequality; where inequality is defined as a measure of the welfare of a society and describes the disparity relative to standards of living across a population (Gitau, 2005). Inequality is also thought of as the differences between an individual or households in terms of opportunities and outcomes, and in Kenya, it shows that, it is manifested in various forms including income from assets, social, political exclusions and inability of certain groups of society to access key social services.

Inequality entails, differences between incomes, differences in access to education, health, land use, land ownership and other welfare enhancing assets and services. Inequality is an important issue in economic development as it can hinder economic growth (Tavneet *et al*, 2008). And to examine income inequality within a given distribution Tavneet used the Gini coefficient which varies from 0 indicating perfect equality to 1 which indicates perfect inequality i.e. one house hold holds all the income of a society. A study by (Jerry *et al*, 2000) suggests that quantifying the impacts of socio economic factors on variability of net farm income is not easy, but on his study he concluded that changes in gross crop income had the largest impact, age was also a key determinant in determining the net crop income variability in a society.

Access to credit is particularly important for farmers given their role as food producers and providers (Quisumbing, 1994). Accessibility to credit allows for access to agricultural inputs thereby impacting farmer's agricultural productivity as indicated by (Saito *et al*, 1994) in his analytical assessment of the gender differentials in agricultural productivity with regard to constraints factors affecting women farmers.

Access to extension services is paramount for agricultural productivities and a shift positively to the rural incomes, because it bridges the gap between technical knowledge and farmers practices (Saito *et al*, 1994). Extension services contributes to the realization of higher levels of agricultural output improving the social position and economic well being of the house holds and a positive boost to the farmers farm income.

The study carried out by Kuria L. M. (2001) in MIS on factors fostering entrepreneurship in agriculture by rice farmers, he came to conclusion that the effects of age, level of education and sources of livelihood were significant in fostering entrepreneurships. A study on technical efficiency in MIS, where two groups of farmers: MMRG and non- MMRG were compared by Kuria Ngige (2008), the conclusion of this study was farmer's characteristics such as Age and education level as well as institutional factors like access to credit and extension facilities are important factors in determining the level of efficiency employed. Achia *et al* (2010) also in his study came to the same conclusion that increase in education level have a positive impact in reducing the possibility that a house hold is poor.

2.2 Overview of hierarchical linear modeling (HLM)

Most of the social, behavioral and even economic data often have a hierarchical structure. A frequently cited example is in education, where students are grouped in classes. Classes are grouped in schools, schools in school districts, etc. We thus have variables describing individuals. HLM have been branded different names such as- multilevel models, mixed-effects models, random-effects models, random coefficient regression models, covariance components models, variance components models- and deal with the analysis of data where observations are nested within groups or higher-order units. In the case of repeated measurement data, the individual serves as the group, with multiple measurements nested within the individual.

Traditionally, fixed parameter linear regression models are used for the analysis of such data, and statistical inference is based on the assumptions of linearity, normality, homoscedasticity, and independence. Ideally, the first of these assumptions should be used for multilevel data. It has been shown by Aitkin & Longford (1986) that the aggregation over individual observations may lead to misleading results in his study on the schools effectiveness study.

Aggregation of, for example, student characteristics over classes facilitate a class analysis, but in the process all individual information is lost. As within-group variation frequently accounts for most of the total variation in the outcome, this loss of information can have an adverse effect on the analysis and lead to distortion of relationships between variables. The alternative, disaggregation, implies the assignation of all class, school, and higher-level characteristics to the individual students. In the process, the assumption of independent observations no longer holds.

Both the aggregation of individual variables to a higher level of observation and the disaggregation of higher order variables to an individual level has been somewhat discredited as (Bryk & Raudenbush, 1992) shows. It has also been pointed out by Holt *et al.* (1980), that serious inferential errors may result from the analysis of complex survey data if it is assumed that the data have been obtained under a simple random sampling scheme.

In hierarchical data, individuals in the same group are also likely to be more similar than individuals in different groups. Due to this, the variations in outcome may be due to differences between groups, and to individual differences within a group. Thus, variance component models, where disturbance may have both a group and an individual component, can be of help in analyzing data of this nature. Within these models, individual residual components are independent, but while group residual components are dependent within groups.

Random regression models have been developed to model educational data where the response variable was a continuous data by (Bock, 1989). In random regression models, however, there is still no possibility of including higher level variables. In order to accommodate both random coefficients and higher order variables, multilevel models should be used.

Multilevel analysis allows characteristics of different groups to be included in models of individual behavior. Most analyses of social sciences data entail the analysis of data with built-in

hierarchies, usually obtained as a sequence of complex sampling methods. Thus, the scope for application of multilevel models is very wide. The formulation of such models and estimation procedures may be seen as an effort to develop a new family of analytical tools that correspond to the classical experimental designs. These models are much more flexible in that they are capable of handling unbalanced data, the analysis of variance-covariance components and the analysis of both continuous and discrete response variables.

As the characteristics of individual groups are incorporated into the multilevel model, the hierarchical structure of the data is taken into account and correct estimates of standard errors are obtained. The exploration of variation between groups, which may be of interest in its own right, is facilitated. Valid tests and confidence intervals can also be constructed and stratification variables used in the sample design can be incorporated into the model.

The use of multilevel models has been hampered in the past by the fact that closed form mathematical formulas to estimate the variance and covariance components have only been available for perfectly balanced designs. The assumption was that each lower-level unit, for example an individual student, was nested within a unique higher-level unit such as a school. In other words, a one to one relationship was assumed to define the nesting of units within groups. Excluded from this were hierarchies in which cross-classification occur, for example where students from multiple neighborhoods may end up going to multiple schools; a situation where students are "cross-classified" by neighborhoods and schools.

To address this situation and allow for the inclusion of explanatory variables for more than one "classification" variable where the coefficients of an individual level model describing the association between individual-level variables and the outcome for groups defined by the "classification" variables, cross-classified random-effect models were developed Goldstein (1995).

An interesting development was in terms of the type of outcome variables considered, where previously, interest was confined to continuous outcome variables, statistical theory has been extended and implemented in statistical software development to appropriately handle binary

outcomes, ordered categorical outcomes, and multi-category nominal scale outcomes within the hierarchical framework.

The next chapter looks at the theoretical model building for a 2 level hierarchical model for clustered data analysis. We shall look at the complex sampling design used for collecting the kind of data structure required for multilevel data and show the theoretical aspect of the different types of model building used in multilevel analysis and there significance in a given study.

CHAPTER THREE:

METHODS

3.1 Introduction

Since we are interested in the description of the multilevel analytical technique to study a continuous response of a clustered data, the chapter begins by looking at the appropriate research design used for clustered survey data and later explains the theory of multilevel model used in the analysis of such a form of data.

3.2 Research design

The study design for the initial survey was a cross-sectional survey conducted in April 2012 at Mwea Irrigation Scheme and the sampling technique used was proportionate stratified technique, where the sample frame (a list of farmers who are beneficiaries of the MIS) was stratified with respect to sections of MIS. Simple Random Sampling (SRS) was performed in each section to achieve a sample size n_j for each section, with a sampling fraction of 0.05

Where $n_j = N_j \times 0.05$ for each stratum j

$N_j = j^{\text{th}}$ Stratum population size, for $j = 1, 2, 3, \dots, N$

$n_j =$ Sample size for the j^{th} stratum (section)

For us to make use of this data, we will assume that each stratum is a cluster of its own. Therefore the MIS baseline survey 2012 has 8 clusters. From these 8 clusters, we performed SRS to come up with 6 clusters for our study and these are Tabere, Mutithi, Thiba, Karama, Mwea and Wamumu as shown in Table 1 below.

Table 1: Sections of the MIS and total number of beneficiaries per section

Section or area	Beneficiaries' number	Sampling number
Tabere	1,100	55
Mwea	928	46
Thiba	868	43
Wamumu	843	42

Section or area	Beneficiaries' number	Sampling number
Karaba	749	37
Nderwa North	727	36
Marura	280	14

3.2.1 Stratified sampling

The research design used for the initial data was stratified sampling where the target population (MIS) was partitioned into geographical units or strata (for this case the MIS sections). The sampling frame was first grouped into stratas and within each stratum SRS was performed. The sampling fraction used for this survey was 0.05, and this implies the design effect was 20.

Sampling fraction = n/N ; where $n = \sum_{j=1}^n n_j$ and $N = \sum_{j=1}^n N_j$; for $j = 1, 2, 3, \dots, N$

Design Effect = N/n where, n =total sample size and N = total population size

The sampling fraction of 0.05 was used in each of these stratum and hence we can say the study design was proportionate stratification. Where,

$$\text{Design effects or weights } Deff = \sigma_{stratification}^2 / \sigma_{SRS}^2$$

$$\sigma_{stratification}^2 = \text{Variance due to stratification}$$

$$\sigma_{SRS}^2 = \text{Variance due to SRS}$$

If weights are not accounted for, then we are likely to have biased estimates.

3.2.2 Cluster sampling

Clustering involves first breaking down a population into a higher level characteristics or unit than the population elements such as a geographical unit.

Clustering can involve one stage (one stage cluster design) or more stages (multi stage cluster design). The higher level cluster is the Primary Sampling Unit (PSU) and lower level units termed as the Secondary Sampling Units (SSU). For our case, the MIS sections form the PSU and the households are the SSU. Individuals in a cluster are assumed to share a common

characteristic and the intra-class correlation coefficient (ICC) or rho is a measure of this homogeneity.

The lack of independence has implications for the design and analysis of survey data. The analysis of cluster designed survey data should take into account the clustered nature of the data. The standard statistical techniques such as Multiple Linear Regression (MLR) and ANOVA are no longer appropriate, unless an aggregate analysis is performed at the cluster level, as they require data to be independent as indicated in the literature chapter. If clustering effects is ignored, p-values will be artificially extreme and confidence intervals will be over-narrow increasing the chances of spuriously significant findings and misleading calculations.

Although an aggregated analysis can be performed at the cluster level using the standard tests, this approach is statistically inefficient because it leads to generalization that is termed as ecological fallacy. Further more, it doesn't allow variation at the lower level to be explored. A more advanced technique have now been developed to analyze level 1 or household level data arising from clustered design and other complex survey designs, which allow the hierarchical nature of the data to be modeled at each level of the data; at the cluster level and the lower level.

3.3 Hierarchical Linear Models

3.3.1 Introduction

Multilevel analysis is a general term referring to statistical methods appropriate for the analysis of data sets comprising several types of unit of analysis (Bliese, 2012).

- i. **Hierarchy:** Organization from detailed to global levels is example normally used; where students are nested in schools, schools nested in there localities e.g. districts and later districts nested in provinces.
- ii. **Level:** Part in a hierarchy, consisting of a collection of units of one type. The most detailed level is level 1.
- iii. **Unit:** Elements belonging to a level.
- iv. **Nesting:** Collection of units belonging to a level.
- v. **Error/residual:** Unexpected variance.

Multilevel data can arise from the study design or a natural hierarchy in the target population, or sometimes both. Multilevel data comes from a data structure in the population that is

hierarchical, with sample data consisting of a multistage sample from this population. Some of the common examples of types of nested data in social research include;

- i. Geographical clustering and group memberships, where individuals are measured at level 1 while at level 2 we may have communities, schools, cities or nations.
- ii. Longitudinal data/panel data, where at level 1 we have multiple measurement of data across time and level 2 individuals or groups.
- iii. Item response theory where at level 1 we have multiple items on test or survey and at level 2 the individuals.

Explanatory Variables characterizing the levels may vary at either of the levels and the response variable Y_{ij} measured at the lowest level. HLM or multilevel modeling techniques is the main model of multilevel analysis and it is designed for hierarchically structured data. It has been documented to have developed rapidly in late 1980's, when the computing methods and resources for this modeling procedure developed. Unlike the standard multiple linear regression (MLR) analysis where single level analysis are performed whether at the group level or individual level, HLM allows us to investigate the relationship of a dependent variable at several levels simultaneously resulting in having more than one residual variance.

A full multilevel regression model assumes that there is a hierarchical data set, with one or more dependent variables that are measured at the lowest level and explanatory variables at all existing levels. Conceptually the model can be viewed as a hierarchical system of regression equations.

3.3.2 Justification in the use of HLM(s)

The HLM is thought to be more powerful compared to the normal standard MLR and OLS regressions techniques. HLM models permits us to; study effects that vary by entity (or groups) and Estimate group level variances and averages measures.

Statistical models that are not hierarchical sometimes ignore nesting structure and therefore report underestimated standard errors. Multilevel techniques assume a General Linear Model framework and can thus perform most types of analyses. Multilevel techniques are thought to examine beyond questions of "Do groups differ?" to ask specific questions of "Why do groups differ?".

Some of the conceptual problems encountered and lead to misinterpretations of results are like;

- i. If we assume that an equation we estimate at the group level also occurs at the individual level, which is to make a cross level inference without allowing the fact that people varies within each group, then this phenomenon is often referred to as an ecological fallacy or the Robinson effects i.e. interpreting aggregated data at the individual level.
- ii. Drawing inferences at higher level from analysis performed at lower level is just as misleading, and this error is termed as atomistic fallacy.
- iii. “Simpson’s paradox” fallacy which refers to the problem that completely erroneous conclusion may be drawn if grouped data, drawn from heterogeneous populations, are collapsed and analyzed as if they came from a single homogeneous population.

3.3.3 A 2 level HLM model for continuous response variable and model Assumptions

Multilevel models represent a comprehensive way to analyze data which are organized in a nesting structure; by this approach it is possible to take into account within-group as well as between-group relations (Snijders, 1999). The dominant approach to analysis of multilevel data employs a type of linear mixed effects model known as the hierarchical linear model. The number of hierarchies will define the kind of multilevel model to use. Such that; for a 2 level hierarchical study we make use of a 2 level HLM model .The response variable is measured at the lowest level 1, but covariates can be measured at any of the levels. The correlation induced by clustering is described by random effects at each level of the hierarchy.

For this study we intend to focus on a 2 level hierarchical linear model for clustered data, where individuals within clusters (sections) are at level 1 and the clusters (sections) of Mwea Irrigation Scheme are at level 2.

3.3.4 Fitting a 2 level hierarchical linear model

For multilevel analysis involving two levels (e.g. individuals nested within clusters/sections/groups), the model can be conceptualized as a two-stage system of equations in which the individual variation within each group is explained by an individual-level equation (level 1), and the variation across groups in the group-specific regression coefficients is explained by a group-level equation (level 2). The case for a normally distributed dependent variable is illustrated below. Let's focus on the case of only one independent variable to give us an illustration of how Hierarchical linear models are constructed and later on build models that extend to include p independent variables.

In multilevel modeling, we are able to construct different kinds of models depending with the kind of research questions that we intend to address or rather depending with the kind of explanatory variables we assume to be random in our targeted population. There are four kinds of hypothesized model constructions, namely;

- i. Null model i.e. a model with no explanatory variables.
- ii. Random intercept models.
- iii. Random slope models.
- iv. Random intercept and random slope models

Let us illustrate the construction of the above kinds of models using one response variable and one explanatory variable and later generalize for a model with p explanatory variables before having a practical application in our current study.

3.3.4.1 Fitting a null model (empty model)

This is a model with no explanatory variables and normally it is the first model to be tested and used to:

- i. Estimate grand mean for dependent variable.
- ii. Estimate variance components at both levels.
- iii. Estimate intra-class coefficients.

The null model facilitates us to test the hypothesis of whether all groups (j units) have the same mean.

Let Y_{ij} be the response variable for the i^{th} individual in group j and x_{ij} is the independent variable for the i^{th} individual in group j where $(i = 1, 2, 3, 4, 5, \dots, n_j)$ and $j = 1, 2, 3, 4, 5, \dots, N$

n_j Represents the total number of observations units /individuals in cluster j

N Represents the total number of clusters size at level 2

To fit the null model, we have the following equations at the 2 levels.

For level 1

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij} \tag{3.1}$$

For level 2

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{3.2}$$

The combined model or marginal model is given as;

$$Y_{ij} = \gamma_{00} + u_{0j} + \varepsilon_{ij} \text{ And } \varepsilon_{ij} \sim N(0, \sigma^2) \quad u_{0j} \sim N(0, \text{var}(u_{0j}))$$

The marginal model is distributed as a normal random variable

$$Y_{ij} \sim N(\gamma_{00}, \text{var}(Y_{ij})) \tag{3.3}$$

$$\begin{aligned} \text{var}(Y_{ij}) &= E[(Y_{ij} - E(Y_{ij}))^2] = E[((\gamma_{00} + u_{0j} + \varepsilon_{ij}) - \gamma_{00})^2] \\ &= E[(u_{0j} + \varepsilon_{ij})^2] = E(u_{0j}^2) + E(\varepsilon_{ij}^2) + 2E(u_{0j}\varepsilon_{ij}) \\ &= \text{var}(u_{0j}) + \text{var}(\varepsilon_{ij}) \text{ and } \text{cov}(u_{0j}, \varepsilon_{ij}) = 0 \end{aligned}$$

$$\text{var}(Y_{ij}) = \tau_{00}^2 + \sigma^2 \tag{3.4}$$

Where;

Y_{ij} is the response variable for the i^{th} individual in the j^{th} group

γ_{00} This is the intercept coefficient and represents the overall group's grand mean or the population means average.

u_{0j} Represents the deviations of the j^{th} group mean from the overall group/cluster average and termed as the group/cluster effects random component or the groups/clusters residuals

ε_{ij} the residual errors between the individuals at level 1

τ_{00}^2 The group level variance as result of the groups/clusters residuals at level 2 or the between group variance

σ^2 Represents the variance at level 1 or the within cluster/group variance

The Intra class correlation (ICC) is the proportion of the total variance that is due to differences among groups (level 2) and can be expressed as;

$$\rho = \tau_{00}^2 / (\tau_{00}^2 + \sigma^2) \quad 3.5$$

$\tau_{00}^2 + \sigma^2$ Represent's the total variance in the response variable(s) for all the individuals in the study.

ρ (Rho) is the interclass or intergroup correlation coefficient (ICC).

3.3.4.2 Fitting a Random intercept model.

For this kind of model, the intercept is assumed to vary among the groups while the explanatory variables are assumed to be fixed.

Let the response variable and the explanatory variable be defined as above.

Equation for Level 1

$$Y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \varepsilon_{ij} \quad \varepsilon_{ij} \sim N(0, \sigma^2) \quad 3.6$$

Equation for Level 2

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_1 &= \gamma_{10} \end{aligned} \quad 3.7$$

The marginal model or combined model is given as;

$$Y_{ij} = \gamma_{00} + u_{0j} + \gamma_{10} x_{ij} + \varepsilon_{ij} \quad 3.8$$

And distributed as

$$Y_{ij} \sim N((\gamma_{00} + \gamma_{10} x_{ij}), \text{var}(Y_{ij})) \quad 3.9$$

$$E(Y_{ij}) = \gamma_{00} + \gamma_{10} x_{ij}$$

And the random effects distributed as $u_{0j} \sim N(0, \tau_{00}^2)$

Where;

γ_{10} The slope or regression coefficient representing the effect for the x_{ij} explanatory variable

The other parameters in the model are defined as above.

$$\begin{aligned} \text{var}(Y_{ij}) &= E[(Y_{ij} - E(Y_{ij}))^2] = E[((\gamma_{00} + \gamma_{10}x_{ij} + u_{0j} + \varepsilon_{ij}) - (\gamma_{00} + \gamma_{10}x_{ij}))^2] \\ &= E[(u_{0j}^2 + 2u_{0j}\varepsilon_{ij} + \varepsilon_{ij}^2)] = E(u_{0j}^2) + E(u_{0j}\varepsilon_{ij}) + E(\varepsilon_{ij}^2) \\ &= \text{var}(u_{0j}) + 2\text{cov}(u_{0j}, \varepsilon_{ij}) + \text{var}(\varepsilon_{ij}) \\ &= \tau_{00}^2 + \sigma^2 \end{aligned} \tag{3.10}$$

Where, $\text{cov}(u_{0j}, \varepsilon_{ij}) = 0$

3.3.4.3 Fitting a random slope model

For the random slope model, the slope varies among the groups while the intercept is assumed to be fixed. Let the response variable and the explanatory variables be defined as above. Then we have the model specification given as;

For level 1

$$Y_{ij} = \beta_0 + \beta_{1j}x_{ij} + \varepsilon_{ij} \tag{3.11}$$

For level 2

$$\begin{aligned} \beta_0 &= \gamma_{00} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \end{aligned} \tag{3.12}$$

The combined model/ marginal model

$$Y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + u_{1j}x_{ij} + \varepsilon_{ij} \tag{3.13}$$

The marginal model is distributed as;

$$Y_{ij} \sim N((\gamma_{00} + \gamma_{10}x_{ij}), \text{var}(Y_{ij}))$$

$$\begin{aligned} \text{And } \text{var}(Y_{ij}) &= E[(Y_{ij} - E(Y_{ij}))^2] = E[((\gamma_{00} + \gamma_{10}x_{ij} + u_{1j}x_{ij} + \varepsilon_{ij}) - E(\gamma_{00} + \gamma_{10}x_{ij}))^2] \\ &= E[(u_{1j}x_{ij} + \varepsilon_{ij})^2] = E[x_{ij}^2u_{1j}^2 + \varepsilon_{ij}^2 + 2x_{ij}u_{1j}\varepsilon_{ij}] = x_{ij}^2E(u_{1j}^2) + E(\varepsilon_{ij}^2) + 2x_{ij}E(u_{1j}\varepsilon_{ij}) \\ &= x_{ij}^2 \text{var}(u_{1j}) + \text{var}(\varepsilon_{ij}) + 2x_{ij} \text{cov}(u_{1j}, \varepsilon_{ij}) \\ &= \tau_{11}^2x_{ij}^2 + \sigma^2 \end{aligned} \tag{3.14}$$

Where, $\text{cov}(u_{1j}, \varepsilon_{ij}) = 0$ and τ_{11}^2 is the level 2 variance as a result of the u_{1j} random effects.

u_{1j} Represent the group deviation from the average slope.

3.3.4.4 Fitting a Random intercept and random slope model

For this model, the intercept and the slope of the model vary among the groups such that we get a distinct regression equation line for each group. The intercept and the slope may be assumed to either co vary or to vary independently.

Let the response and the explanatory variables be defined as above. Then the model equations are as follows;

For Level 1

$$Y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij} \quad \varepsilon_{ij} \sim N(0, \sigma^2) \quad 3.15$$

For level 2

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad 3.16$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

Where the random effects are assumed to be distributed as follows;

$E(u_{0j}) \sim N(0, \tau_{00}^2)$ and $E(u_{1j}) \sim N(0, \tau_{11}^2)$ and $\text{cov}(u_{0j}, u_{1j}) \neq 0$

The marginal model is given as;

$$Y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + u_{0j} + u_{1j}x_{ij} + \varepsilon_{ij} \quad 3.17$$

And

$$E(Y_{ij}) = \gamma_{00} + \gamma_{10}x_{ij}$$

Hence distributed as; $Y_{ij} \sim N((\gamma_{00} + \gamma_{10}x_{ij}), \text{var}(Y_{ij}))$

Where,

$$\begin{aligned} \text{var}(Y_{ij}) &= E[((Y_{ij} - E(Y_{ij}))^2] = E[(\gamma_{00} + \gamma_{10} + u_{0j} + u_{1j}x_{ij} + \varepsilon_{ij}) - E(\gamma_{00} + \gamma_{10}x_{ij})]^2] \\ &= E[(u_{0j} + u_{1j}x_{ij} + \varepsilon_{ij}')^2] \end{aligned}$$

$$\begin{aligned}
&= E(u_{0j}^2) + x_{1j}^2 E(u_{1j}^2) + E(\varepsilon_{ij}^2) + 2x_{1j} E(u_{0j}u_{1j}) + 2E(u_{0j}\varepsilon_{ij}) + x_{1j} E(u_{1j}\varepsilon_{ij}) \\
&= \text{var}(u_{0j}) + x_{1j}^2 \text{var}(u_{1j}) + \text{var}(\varepsilon_{ij}) + 2x_{1j} \text{cov}(u_{0j}, u_{1j}) + 2\text{cov}(u_{0j}, \varepsilon_{ij}) + x_{1j} 2\text{cov}(u_{1j}, \varepsilon_{ij}) \\
&= \tau_{00}^2 + x_{1j}^2 \tau_{11}^2 + 2x_{1j} \tau_{10} + \sigma^2
\end{aligned}$$

3.18

And, $\text{cov}(u_{0j}, \varepsilon_{ij}) = 0$, $\text{cov}(u_{0j}, u_{1j}) \neq 0$

Hence

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00}^2 & \tau_{01} \\ \tau_{01} & \tau_{11}^2 \end{pmatrix} \text{ And in matrix notations } U_j \sim N_2(0, T)$$

In matrix form, the above model can be written as,

$$Y_{ij} = \gamma_{00} + \gamma_{10}x_{1j} + u_{0j} + u_{1j}x_{1j} + \varepsilon_{ij}$$

$$\begin{pmatrix} Y_{1j} \\ Y_{2j} \\ Y_{3j} \\ \vdots \\ Y_{njj} \end{pmatrix} = \begin{pmatrix} 1 & x_{1j} \\ 1 & x_{2j} \\ 1 & x_{3j} \\ \vdots & \vdots \\ 1 & x_{njj} \end{pmatrix} \begin{pmatrix} \gamma_{00} \\ \gamma_{10} \end{pmatrix} + \begin{pmatrix} 1 & x_{1j} \\ 1 & x_{2j} \\ 1 & x_{3j} \\ \vdots & \vdots \\ 1 & x_{njj} \end{pmatrix} \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1j} \\ \varepsilon_{2j} \\ \varepsilon_{3j} \\ \vdots \\ \varepsilon_{njj} \end{pmatrix}$$

3.19

$$Y_{ij} = X\beta + ZU_j + \varepsilon_j \text{ For this case } p = q = 2$$

$$\text{And } \text{var}(Y_{ij}) = ZTZ' + \sigma^2$$

3.20

The distribution of the marginal model is given as;

$$Y_{ij} \sim N(X\beta, ZTZ' + \sigma^2)$$

3.21

Where $E(Y_{ij}) = X\beta$ and $\text{var}(Y_{ij}) = ZTZ' + \sigma^2$ for this case $p = q = 2$

3.3.3.4 The compressed case for the hypothesized models

For the Random intercept and Random slope model with p fixed explanatory variables, q random effects and one response variable we have for;

Level 1

$$Y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + \dots + \beta_{pj}x_{pij} + \varepsilon_{ij}$$

$$= \beta_{0j} + \sum_{k=1}^p \beta_{kj}x_{kij} + \varepsilon_{ij} \quad 3.22$$

Level 2

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + u_{2j}$$

⋮

⋮

⋮

$$\beta_{pj} = \gamma_{p0} + u_{pj} \quad 3.23$$

We can write the above equations as $\beta_{kj} = \gamma_{k0} + u_{kj}$ where $k = 1, 2, 3, \dots, p$

The combined model is provided as;

$$Y_{ij} = \gamma_{00} + \sum_{k=1}^p \gamma_{k0}x_{kij} + u_{0j} + \sum_{k=1}^p u_{kj}x_{kij} + \varepsilon_{ij} \quad \text{Where } p \geq q \quad 3.24$$

The distribution of the marginal model is given as;

$$Y_{ij} \sim N(E(Y_{ij}), \text{var}(Y_{ij})) \quad \text{Where } E(Y_{ij}) = X\beta = \gamma_{00} + \sum_{k=1}^p \gamma_{kj}x_{kij}$$

And

$$\text{var}(Y_{ij}) = \tau_{00}^2 + 2 \sum_{k=1}^q \tau_{k0}x_{kij} + \sum_{k=1}^q \tau_k^2 x_{kij}^2 + 2 \sum_{k<l} \tau_{kl}x_{kij}x_{lij} + \sigma^2 \quad 3.25$$

The variance of the response variable is a function of the explanatory variables.

The case for the random slope model with q random effects components and fixed intercept, we have;

$$Y_{ij} = \gamma_{00} + \sum_{k=1}^p \gamma_{k0}x_{kij} + \sum_{r=1}^q u_{rj}x_{rj} + \varepsilon_{ij} \quad \text{Where } p \geq q \quad 3.26$$

The distribution of the marginal model is given as;

$$Y_{ij} \sim N(E(Y_{ij}), \text{var}(Y_{ij}))$$

Where;

$$\text{var}(Y_{ij}) = \sum_{k=1}^q \tau_k^2 x_{kij}^2 + \sigma^2 \quad 3.27$$

For a model with p explanatory variables, we make use of the Laird and Ware's matrix formulation (Laird and Ware, 1982) which illustrates an appropriate way in defining a 2 level Hierarchical linear model, and structured as follows:

$$\begin{pmatrix} Y_{1j} \\ Y_{2j} \\ Y_{3j} \\ \vdots \\ Y_{njj} \end{pmatrix} = \begin{pmatrix} 1x_{11j} x_{11j} x_{11j} \dots x_{p1j} \\ 1x_{12j} x_{12j} x_{12j} \dots x_{p2j} \\ 1x_{13j} x_{13j} x_{13j} \dots x_{p3j} \\ \dots \\ 1x_{1njj} x_{1njj} x_{1njj} \dots x_{pnjj} \end{pmatrix} \begin{pmatrix} \gamma_{00} \\ \gamma_{10} \\ \gamma_{20} \\ \vdots \\ \gamma_{p0} \end{pmatrix} + \begin{pmatrix} 1x_{11j} x_{11j} x_{11j} \dots x_{q1j} \\ 1x_{12j} x_{12j} x_{12j} \dots x_{q2j} \\ 1x_{13j} x_{13j} x_{13j} \dots x_{q3j} \\ \dots \\ 1x_{1njj} x_{1njj} x_{1njj} \dots x_{qnjj} \end{pmatrix} \begin{pmatrix} u_{1j} \\ u_{2j} \\ u_{3j} \\ \vdots \\ u_{njj} \end{pmatrix} + \begin{pmatrix} \epsilon_{1j} \\ \epsilon_{2j} \\ \epsilon_{3j} \\ \vdots \\ \epsilon_{njj} \end{pmatrix} \quad 3.28$$

$$\underline{Y}_{-j} = \underline{X}_{-j} \underline{\beta} + \underline{Z}_{-j} \underline{U}_{-j} + \underline{\epsilon}_{-j}$$

Where,

$\underline{Y}_{-j} = n_j \times 1$ Is a random vector whose components are the Y_{ij} ($i = 1, 2, 3, 4, 5, \dots, n_j$)

$\underline{\beta} = p \times 1$ The vector for the regression coefficients of the explanatory variables

$\underline{X}_{-j} = n_j \times p$ The design matrix for the explanatory variables and whose components are the x_{ij} 's

$\underline{Z}_{-j} = n_j \times q$ The design matrix for the random effects variables and whose components are

x_{ij} 's Those are assumed to be random or vary across the group level or level 2.

$\underline{U}_{-j} = q \times 1$ The vector for group's random effects or group residuals

$\underline{\epsilon}_{-j} = n_j \times 1$ The vector for the random error or individual residuals

3.3.5 Model Assumptions

Just like any other statistical analytical technique, HLM has model assumptions that are to be looked at before data analysis and concise interpretation procedure is performed. Some of these assumptions include;

- a) The continuous response variables are assumed to be normally distributed.
- b) The random part of the model assumptions are;
 - i. The mean vectors

$$E(U_j) = 0 \text{ And } E(\varepsilon_{ij}) = 0$$

- ii. Variance matrices $Var(\varepsilon_{ij}) = \sigma^2 I_{(n_j)}$ and $Var(U_j) = T$

- iii. Covariance matrices $cov(\varepsilon_{ij}, u_{0j}) = 0$

- iv. and $cov(u_{kj}, u_{k'j}) \neq 0$

The residuals and the random effects are assumed to be distributed according to a multivariate normal distribution i.e.

$$\varepsilon_{ij} \sim N_{n_j}(0, \sigma^2 I_{(n_j)})$$

$$U_j \sim N_q(0, T), \text{ for } j = 1, 2, 3, 4, 5, \dots, N$$

The parameters τ_{hm} in the variance covariance matrix T , where $(hm = 0, 1, 2, \dots, q - 1)$ and the residual variance (level1) $= \sigma^2$, are termed as the variance random components.

As cluster effect subscript j is present for Y_j and X_j , meaning cluster sample sizes can vary.

The design matrix X_j can include; Covariates measured at level 1 (individual level), Covariates measured at cluster level i.e. level 2 and Cross interaction of covariates at both levels.

$p =$ The total number of covariates/explanatory independent variables and the random component;

- i. U_j distinguishes the HLM from the standard (fixed effects) MLR
- ii. U_j represents effects of level 1 clustering (one for every cluster)

3.3.6 Hypothesis Testing (parameter testing's)

We can perform two types of hypothesis testing using the multilevel model analysis, and these are the single parameter tests e.g. significance of single predictor variable in multiple equations and the multiple parameter tests (Differences across models) e.g. significance of multiple predictors' variable in multiple equations.

The choice depends on the particular hypothesis we want to test and it's advisable to start with a single parameter tests.

3.3.6 Testing the fit of the model

The Likelihood Ratio Test (LRT) is the preferred test statistic used to test the effectiveness of a multilevel model, although other test statistics such as the Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC) are used. When two model formulations are to be tested such as model 1 nested within model 2; then, the model with the lowest value for AIC or BIC is always assumed to be the best model. For LRT test statistics we make use of the log likelihood values, in that;

Suppose Model 1 is nested within Model 2; then the difference between the model's Log likelihoods is assumed to have a chi-square distribution. Hence,

$$2 \times \log(L_{full\ model} / L_{reduced\ model}) = 2 \times (\log L_{full\ model} - \log L_{reduced\ model}) \sim \chi_q^2 \text{ Having } q \text{ degrees of freedom}$$

Where q = number of additional parameters in Model 2

$-2 \log L$ Is called the deviance denoted by D_i where i is the number of the model fit (the higher the deviance the poorer the model fit)

$$D_1 - D_2 \sim \chi_q^2$$

To test the additional parameter(s) in Model 2 jointly equal to zero with respect to the null hypothesis.

LR tests with halved p-values (akin to one-tailed p-values), for tests of variance and covariance parameters is recommended.

3.3.7 Estimation procedure

Multilevel analysis or HLM makes use of the Full Maximum Likelihood Estimation (FMLE) and Restricted Maximum Likelihood estimation (REML) methods to estimate the fixed effects and

the variance components parameters. Each of the mentioned estimation method has its own advantages although they both produce identical fixed effects estimates. FMLE takes into account the degree of freedoms from the fixed effects and hence produces variance components estimates that are less biased, in addition to the ability to handle unbalanced data. REML is most preferred for small samples with balanced data because it is unbiased, but for large samples the difference between estimates are negligible. One disadvantage of REML is that the likelihood ratio test can not be used to compare models with different fixed effects specifications. FMLE can be used to test for differences between models that differ only in their fixed part and REML can be used to test difference between models that only differ in their random part (the variance components).When computing FMLE it takes longer than REML, because the maximum likelihood estimates uses an iterative procedure in estimating the parameter estimates.

Both of these methods require distributional assumptions, and usually, it is assumed that the residual component ε_{ij} as well as the group random effects components U_j exhibits a Multivariate Normal distribution.

3.4 Measurement instruments

There are quite a number of different kinds of measurements instruments used in survey for data collection and for this survey the instrument used was a structured questionnaire. The questionnaire was filled by the household head who was guided by an enumerator.NIB conducted the survey in April 2012 at MIS.

3.5 Statistical software

The computing technology has developed to include statistical packages that are currently used for the analysis of multilevel models and most of these analytical procedures developed in early 1980's when the multilevel techniques started to gain momentum in its use. Such packages include the Spss, HLM, MLwin, Stata, SAS, R, and others. The statistical package used in this study is the R software because it has a strong graphical tools kit for plotting and easy syntax commands to run the multilevel models. Stata was also used in the initial stages of data analysis to confirm and assert the results output from R.

The R software makes use of the *lme4* library to run the *lmer* function that is used in multilevel models analysis while the Stata package makes use of the *xtmixed* commands.

3.6 Hypothesized models in the study

Before the construction of the hypothesized model in this study, let us specify the variables to be used in the models constructions.

- age_{ij}* This is a measure of age for the *ith* individual in the *jth* section
- educ_{ij}* a measure of level of education for the *ith* individual in the *jth* section
- Ms_{ij}* a measure of the marital status for the *ith* individual in the *jth* section
- gen_{ij}* a measure of the type of gender for the *ith* individual in the *jth* section
- cred_{ij}* a measure of availability of credit facility for the *ith* individual in the *jth* section
- extserv_{ij}* a measure o the availability of extension services for the *ith* individual in the *jth* section

3.6.1 The null model or unconditional model

The null model or empty model is a model with no explanatory variable. For HLM models, the null model is a model with a random intercept component i.e. the intercept is allowed to vary across groups/clusters/sections and this results in 2 random effects components; one at the level 1 and the other at the level 2.

Model specification:

Level-1 model:

$$Y_{ij} = \beta_{0j} + e_{ij} \tag{3.30}$$

$$e_{ij} \sim N(0, \sigma^2) \text{ For } i = 1, 2, 3, 4, 5, \dots, n, \quad j = 1, 2, 3, 4, 5, 6$$

Level-2 model:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{3.31}$$

Combined model:

$$Y_{ij} = \gamma_{00} + u_{0j} + \varepsilon_{ij} \quad 3.32$$

Where,

Y_{ij} This is the gross family income for the i^{th} individual in the j^{th} section

γ_{00} = the population average gross family income.

u_{0j} = this is the random effect associated with J^{th} section and measures the deviation of sectional mean gross family income from the population average gross family income

ε_{ij} = this is the residual error at the individual level

$$Var(Y_{ij}) = Var(\gamma_{00} + u_{0j} + \varepsilon_{ij}) = \tau_{00}^2 + \sigma^2 \quad 3.34$$

$Var(u_{0j}) = \tau_{00}^2$ Represent's the sectional variance in gross family income in MIS.

σ^2 Represent's the variance in gross family income between the individuals in MIS.

3.6.2 A model incorporating the demographic factors as fixed effects.

A model incorporating the demographic factors of gender, age, education level and marital status as fixed effects parameters.

Model specifications;

Level 1 model

$$Y_{ij} = \beta_{0j} + \beta_1 Age_{ij} + \beta_2 Educ_{ij} + \beta_3 Ms_{ij} + \beta_4 gen_{ij} + \varepsilon_{ij} \quad 3.35$$

$$\varepsilon_{ij} \sim N(0, \sigma^2) \quad \text{For } i = 1, 2, 3, \dots, nj \text{ and } j = 1, 2, 3, 4, 5, 6$$

Level 2 model

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_1 &= \gamma_{10} \\ \beta_2 &= \gamma_{20} \\ \beta_3 &= \gamma_{30} \\ \beta_4 &= \gamma_{40} \end{aligned} \quad 3.36$$

Combined model

$$Y_{ij} = \gamma_{00} + \gamma_{10} Age_{ij} + \gamma_{20} Educ_{ij} + \gamma_{30} Ms_{ij} + \gamma_{40} gen + u_{0j} + \varepsilon_{ij} \quad 3.37$$

Where,

γ_{00} = overall mean for gross family income in all the sections.

β_1 =Regression coefficient for age and measures the fixed effect of age in the model.

β_2 =Regression coefficient for level of education and measures the fixed effect of education in the model.

β_3 =Regression coefficient for marital status and measures the fixed effect of marital status in the model.

β_4 =Regression coefficient for gender and measures the fixed effect of gender in the model

u_{0j} Is the random component for the intercept and measures the sectional effects.

3.6.3 A model incorporating the institutional factors as fixed effects.

A model incorporating the institutional factors; i.e. access to credit and access to extension services as fixed effects parameters.

Model specifications;

Level 1 model

$$Y_{ij} = \beta_{0j} + \beta_1 cred_{ij} + \beta_2 extserv_{ij} + \varepsilon_{ij} \quad 3.38$$

Level 2 model

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_1 = \gamma_{10}$$

$$\beta_2 = \gamma_{20} \quad 3.39$$

Combined model

$$Y_{ij} = \gamma_{00} + \gamma_{10} cred_{ij} + \gamma_{20} extserv_{ij} + u_{0j} + \varepsilon_{ij},$$

$$\varepsilon_{ij} \sim N(0, \sigma^2), u_{0j} \sim N(0, \tau_{00}^2) \quad 3.40$$

3.6.4 A model having demographic factors as random effects.

This model assumes all the factors comprising demographic factors in the study are random and hence we have different slopes for each of the factors.

Model specifications;

Level 1 model

$$Y_{ij} = \beta_{0j} + \beta_{1j} Age_{ij} + \beta_{2j} Educ_{ij} + \beta_{3j} Ms_{ij} + \beta_{4j} gen_{ij} + \varepsilon_{ij} \quad 5.41$$

Level 2 model

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \\ \beta_{2j} &= \gamma_{20} + u_{2j} \\ \beta_{3j} &= \gamma_{30} + u_{3j} \\ \beta_{4j} &= \gamma_{40} + u_{4j} \end{aligned} \quad 5.42$$

Combined model

$$Y_{ij} = \gamma_{00} + \gamma_{10} Age_{ij} + \gamma_{20} Educ_{ij} + \gamma_{30} Ms_{ij} + \gamma_{40} gen_{ij} + u_{0j} + u_{1j} Age_{ij} + u_{2j} Educ_{ij} + u_{3j} Ms_{ij} + u_{4j} gen_{ij} + \varepsilon_{ij} \quad 5.43$$

In general;

The gamma symbol γ_{i0} represents regression coefficients' for the respective factors considered in the study and the u_{ij} represent's the random effects for the randomized factors under study.

And $i = 0, 1, 2, 3, 4$

The other parameters are defined as above.

CHAPTER FOUR:

RESULTS

4.1 Data

The data utilized for this study, was collected through a cross sectional survey carried out by NIB in April 2012 in the study area of Mwea Irrigation scheme.

4.1.2 Data description and variable specifications

The variables used for this study are;

1. Explanatory variables/Independent variables:

- i. Gender of the Household Head- a categorical variable with 2 levels
0=male, 1=females
- ii. Level of education for the Household Head- a categorical variable with 2 levels.
- iii. Marital Status of the household head – a categorical variable with 2 level
- iv. Credit facility – a categorical variable with 2 levels;
 - a) 0 = Yes –for those Household Heads that accessed credit facility in the year 2010/2011
 - b) 1 = No - for those Household Heads that did not access credit facility in the year 2010/2011.
- v. Age of the Household Head- a continuous variable and centered around the Grand Mean.
- vi. Extension services - a categorical variable with 2 levels
 - a) 0 = Yes for those Household Heads that accessed or sorted the help of extension agricultural officers.
 - b) 1 = No for those Household Heads that did not access help from agricultural officers.

2. Response variable/Dependent variable:

- i. Gross family income for the Household head in the year 2010/2011 and calculated as follows;

$$\text{Gross family Income} = \text{Farm Income} + \text{Agricultural labor income} + \text{non farm income}$$

Where;

- i. Farm income was calculated as the total crop production (kgs) of various types of crops harvested during short and long rain seasons using irrigation and non irrigation means multiplied by the unit value to sell per kg for each type of crop.
- ii. Agricultural labor income was calculated as Daily wage per day multiplied by the total number of days worked by the Household in farming activities per year.
- iii. Non farm income was calculated as a sum of composite of several items, and this included;
 - a) Salary from other occupations, such as government employee, pension's e.t.c.
 - b) Wages from casual work.
 - c) Earnings from sale of fuel wood and forest products.
 - d) Earnings from sale of livestock and poultry products.
 - e) Earnings from sale of fodder products.
 - f) Earnings from tree crops; own trees and other trees.
 - g) Earnings from home garden.
 - h) Earnings from subsidiary business after subtracting cost for the business.
 - i) Receipt of gifts and remittances from relatives and others.
 - j) Income from lending charges for farm machinery equipment and work animals after deducting the cost for the business or service offered.
 - k) Income from leased land.
 - l) Interest earned from money loan to other persons and bank deposits.
 - m) Others source.

4.1.3 Data preparation

The following guidelines need to be adhered to before running the analysis using the various statistical packages available for multilevel analysis. In order for the result interpretations to be meaningful and easy, there is need to centre continuous explanatory variables and to code categorical variables with reference codes. A variable can be centered using the Grand mean or the group means or we can use the variables Z- scores.

The variable age is a continuous variable and hence before inclusion of this variable into the analysis, we need to centre it using the Grand Mean, and this is provided as follows;

Let X_{ij} be the age of the i^{th} individual in the j^{th} section, then;

$$\text{Centered Age} = X_{ij} - \bar{X}$$

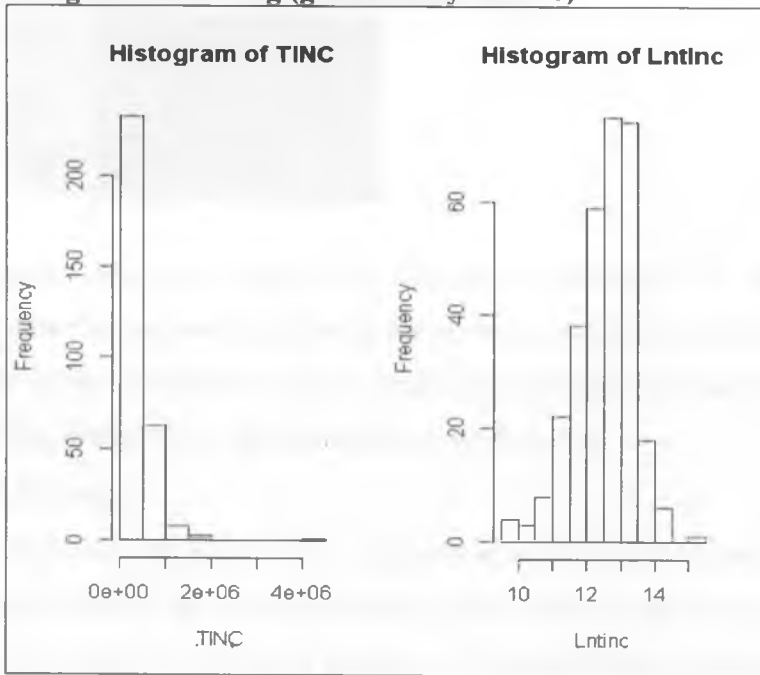
Where; \bar{X} =Grand mean of the Age of the Households in MIS scheme.

For easier result interpretations, the categorical variables were coded with 0 being the reference category. Hence the categorical variables in the study were coded as follows; level of education (0=primary 1=past primary), Marital status (0=married, 1=Not married) Credit access (0=yes, 1=no), Extension services (0=yes, 1=No) gender (0=male, 1= female). Before running the analysis, the data needs to be arranged or sorted by group level variable in an ascending order or descending order.

4.2 Analysis and Results Interpretations

In analytical assignments, it is advisable to perform Exploratory Data Analysis to get an insight into how the data behaves. We will make use of graphs and frequency tabulations to understand our dataset.

Figure 1: A histogram for the log (gross family income)



TINC=Gross family Total Income tends to be significantly positively skewed and hence to make adjustments for it be linear, we performed a logarithmic transformation to make the income distribution symmetrical as shown in the histogram of Lntinc, where Lntinc is the log transformation of the Gross family total income variable.

Figure 2: Frequency table showing the percentage of the house holds who accessed credit in all the MIS sections.

Valid	yes	61	19.6
	no	250	80.4
	Total	311	100.0

Figure 3: Frequency table for the level of education

primary	226	72.7
school		
past primary	85	27.3
Total	311	100.0

Figure 4: Frequency table for the extension services

yes	100	32.2
no	211	67.8
Total	311	100.0

The above figures tabulations' shows the frequency tabulations of the percentage of the households in MIS that respondent to the levels of the given factors under study. We are able to note that; 27.3% of the respondents had education past primary level and 67.8% never accessed extensions services while 80.4% did not access credit facilities.

4.2.1 Model fitting

The models to be fitted with respect to the objectives of the study comprise the Null model, the Random intercept models, the Random Slopes models and the Random Intercept and Random Slope models. We intend to use the R software although in some instances the Stata software may be used.

4.2.1.1 Fitting the null model

Table 2: Results for the Null Model.

Random effects			
Group	variance	Standard deviation	
Sections	0.052703	0.22957	
Residual	0.791860	0.88987	
Fixed effects			
	Estimate	Std error	t-value
intercept	12.5581	0.1072	117.2

The above result shows a multilevel model that allows sections effects on gross income distribution, but without the explanatory variables. The estimated gross family income for each section is provided for equation 4.1.

$$\text{Log (gross income)}_{ij} = 12.5581 + u_{0j} \quad 4.1$$

Where;

u_{0j} Are the sections effects and corresponds to the deviations of section j 's mean gross income from the overall aggregate mean gross income of all the sections.

The u_{0j} 's and there respective standard errors are provided in Table 3 below.

Table 3: Sectional Effects and their respective standard errors

Sections	Id no.	U_{0j} value	U_{0j} standard errors
Karaba	1	-0.16113168	0.12221367
Mutithi	2	-0.31330181	0.08810756
Mwea	3	0.18668918	0.11392210
Tabere	4	0.03440798	0.10634983
Thiba	5	0.05934552	0.11682982
Wamumu	6	0.19399081	0.11784968

$\gamma_{00} = 12.5581$ it is a regression coefficient representing the average gross income among the MIS sections which is Kshs. 284,389.5

The mean gross family income for section j is estimated as $12.5581 + u_{0j}$ where u_{0j} are the sections residuals which are estimated shown above. A section with $u_{0j} > 0$ has a mean gross family income that is higher than the average gross family income in MIS, while $u_{0j} < 0$ for a below average for sections j .

According to the objective of the research in determining whether there is variability among the sections of MIS, we need to focus on the variance components and their significance to establish whether these kind of dataset warrant us to use the multilevel model or an ordinary linear square regression model (OLS). Partitioning the variance from the analysis; the sectional variance (level 2) is estimated to be $\tau_{00}^2 = 0.052703$ (0.22957) and the variance at the individual level (level 1) estimated to be $\sigma^2 = 0.791860$ (0.88987) and the total variance estimated to be $\text{var}(Y_{ij}) = 0.052703 + 0.791860 = 0.844563$ (0.9190). The parenthesis contains the standard deviations of the respective variances.

When using the Stata Software, both variances show Confidence Intervals that are significant. The ICC = 0.06240269 and this tells us the correlation of the individuals observations in their gross family income within a section or cluster. From the ICC figure we note that 6.24% of gross family income can be attributed to differences between the sections. We should note that when the ICC approaches 0 then there are no sectional effects and hence performing a normal standard OLS preferred. And when ICC equals to 1 then there is no variance at the individual level implying that all individuals have the same gross income across the sections.

Testing for section effects

We make use of equation 4.2 to test the section effects.

$$H_0 : u_{0j} = 0 \text{ Against } H_a : u_{0j} \neq 0$$

4.2

To test section effects, we use the Likelihood Ratio Test (LRT) comparing the null multilevel model with a null single level model. To fit the null single level model, we remove the random section effect in equation 3.32 and fit equation 4.3.

$$\text{Log}(\text{income})_{ij} = \beta_0 + \varepsilon_{ij}$$

4.3

Table 4: Results for the null single level model

Coefficients:	Estimate	Std. Error	t value	p-value
(Intercept)	12.51892	0.05203	240.6	0.0000
Residual standard error: 0.9175 on 310 degrees of freedom				

Obtaining the Log likelihood values for the models and calculating the LRT test statistic i.e.

$LR = 2 * (-410.2160 - -414.0036) = 7.5752$ and this value is greater than $\chi^2_{(1)} = 3.85$. We find that the random section effect is significant and hence a multilevel analysis warranted.

4.2.1.2 Modeling demographic factors

Before modeling the factors that constitute demographical factors in the study, let us examine the relationship for each of these factors to the gross family income.

4.2.1.2.1 Effects of Gender

Table 5: Multilevel Model with Gender as the Explanatory Variable

Models	Null model or Empty model	Fixed intercept Random slope model	Random intercept fixed slope model.	Random intercept random slope model
	MODEL(1)	MODEL(2)	MODEL(3)	MODEL(4)
Fixed parts				
intercept	12.5581(0.1072)	12.5699	12.57305(0.11235)	12.58180 (0.12235)
gender	-	-	-0.06557(0.12419)	- 0.06306(0.1419)
Random part				
τ_{00}^2	0.052703	0.053125	0.054829	0.068732

τ_{10}				-0.04461347
τ_{11}^2		1.255×10^{-07}		0.0289558
σ^2	0.791860	0.79135	0.793342	0.789334
ICC	0.0624027			

Model specification for the Table 5 above is as follows.

Model 1 $Y_{ij} = \beta_{0j} + \varepsilon_{ij}$

Model 2 $Y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}$

Model 3 $Y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \varepsilon_{ij}$

Model 4 $Y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \varepsilon_{ij}$

4.5

Model 3 in Table 5 is what is termed as a Random Intercept with fixed explanatory variables model, where the fixed explanatory variable for this case is gender. We are allowing a linear gender effect and the resulting equation for estimating the average fitted regression line for gross family income in each of the sections accounting for gender effect is provided for by equation 4.6.

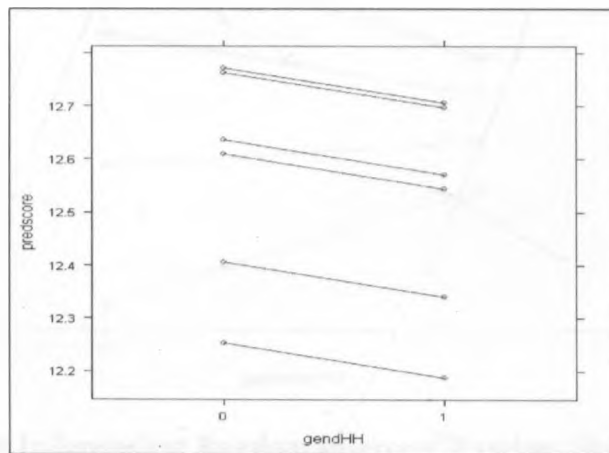
$$\text{Log}(\text{income})_{ij} = 12.57305 - 0.06557 \text{gend}_{ij} + u_{0j}$$

4.6

Figure 5: Plot of estimated income against gender.

The fitted regression lines are depicted in figure 5 below; for a given section will differ from this average line by an amount u_{0j} for section j and whose section variance is given by $\tau_{00}^2 = 0.054829$.

The variance between the individuals gross family income is given as $\sigma^2 = 0.793342$. The slope of the sections regression lines are assumed to be fixed at -0.06557 and non significant implying that, on average there is no difference in Gross family income between males and females within the sections. After accounting for the gender effect, the proportion of unexplained variance that is due to the differences in Gross family income between sections has marginally increased by less than 1%.



When we fit a random intercept random slope model (model 4 in Table 5), where we let both the intercepts and the slope of the estimated regression line to co vary randomly across the sections, the resulting variance components are; variance of the sectional residuals in Gross family income is $\tau_{00}^2 = 0.068732$, the variance as a result of the deviation's from the average slope effect i.e. (u_{1j}) is $\tau_{11}^2 = 0.0289558$, and the covariance $\tau_{10} = -0.04461347$ and finally the variation in Gross family income between individuals $\sigma^2 = 0.789334$. In the random intercept random slope model the average gross family income is $\gamma_{00} = 12.5818$ and significant, while the random slope coefficient $\gamma_{10} = -0.063606$ and non significant, implying the difference in Gross family income between the males and females across the sections is non significant.

Figure 5: A plot of the estimated income against gender for independent random intercept random slope model

Although there is non significance difference in Gross family income between the genders, when we fit an independent random intercept random slope model (where we allow the intercept and the slope to vary independently), the study noted the variation in gross family income distribution is highest for males at 7.9% across the sections with respect to that for females at 1% across the MIS sections as depicted in Table 6 and figure 5 below.

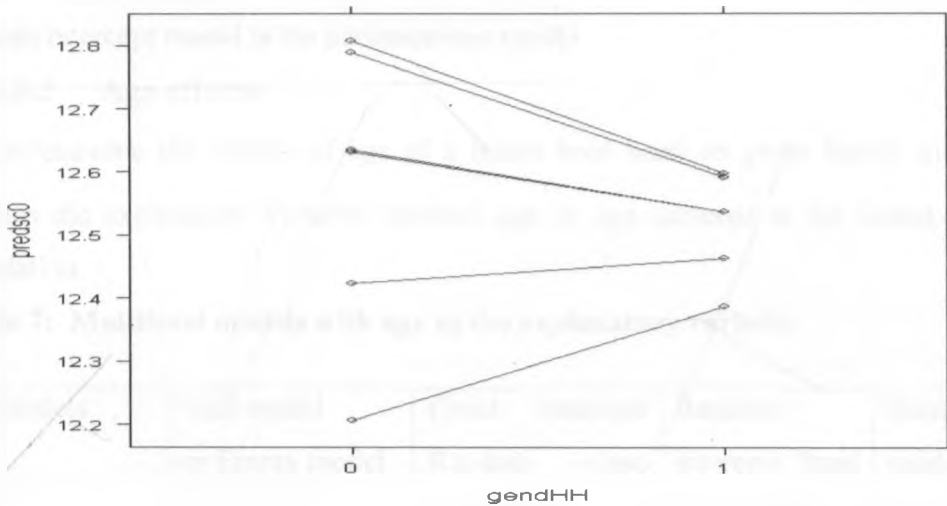


Table 6: Results for the Independent Random Intercept Random Slope Model

Random effects:				
Groups	Name	Variance	Std.Dev.	
section	(Intercept)	1.8383e-16	1.3559e-08	
section	male	6.8732e-02	2.6217e-01	
	female	8.4633e-03	9.1996e-02	
Residual		7.8933e-01	8.8844e-01	
Fixed effects:				
		Estimate	Std. Error	t value
	(Intercept)	12.58180	0.12235	102.83
	gendHH1	-0.06306	0.14186	-0.44

To test whether gender effect varies across the MIS sections, we make use of equation 4.7.

$$H_0 : \tau_{00}^2 = \tau_{11}^2 = 0 \text{ Against } H_a : \tau_{00}^2 \text{ and } \tau_{11}^2 \neq 0 \quad 4.7$$

To test the above hypothesis, we will use the likelihood ratio statistics given as;

LR=2*(-410.5 --411.2) = 1.4, this is less than $\chi_1^2=3.85$ and hence the gender random effects across the sections is non significant. Alternatively we may use the AIC and the BIC values where the model with lowest values for the above statistics is the best model. And for this case, the model with fixed effects for gender without the random component is the best, although the random intercept model is the parsimonious model

4.2.1.2.2 Age effects

Let us examine the effects of age of a house hold head on gross family income. Where x_{ij} denotes the explanatory variable centered age or age centered at the Grand mean age of the population.

Table 7: Multilevel models with age as the explanatory variable

Models	Null model or Empty model.	Fixed intercept Random slope model	Random intercept fixed slope model	Random intercept random slope model
	MODEL(1)	MODEL(2)	MODEL(3)	MODEL(4)
Fixed parts				
intercept	12.5581(0.1072)	12.5699	12.558(0.109)	12.5576(0.1094)
Centered age	-	-	0.00311(0.00311)	0.00311(0.00311)
Random part				
τ_{00}^2	0.052703	0.053125	0.055237	0.055453
τ_{10}	-	0.0002582046	-	-0.0001733589
τ_{11}^2	-	1.255×10^{-07}	-	5.4196×10^{-07}
σ^2	0.791860	0.79135	0.791324	0.79114
ICC	0.0624027			

Models specifications for table 7

Model 1 $Y_{ij} = \beta_{0j} + \varepsilon_{ij}$

Model 2 $Y_{ij} = \beta_0 + \beta_{1j}x_{ij} + \varepsilon_{ij}$

Model 3 $Y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij}$

4.8

Model 4 $Y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij}$

Looking at the fixed part of all the models in table 7, we are able to note that the intercept γ_{00} is significant in all the models unlike the centered age effect γ_{10} which is non significant in all the models.

The issues we may want to address are;

- i. whether gross family income is related to age
- ii. whether the relationship varies systematically or significantly by sector
- iii. whether the age effect varies randomly across sectors

The tabulation below shows the test for the above 3 questions, and we make use of the LRT test to verify this.

Table 8: Results for Models Hypothesis testing

	Df	AIC	BIC	logLik	Chisq	Df	Pr(>Chisq)
MODEL (1)	3	823.80	835.02	-408.90			
MODEL (2)	4	824.81	839.77	-408.40	0.9959	1	0.3183
MODEL (3)	5	827.67	846.37	-408.83	0.0000	1	1.0000
MODEL (4)	6	828.76	851.20	-408.38	0.9055	1	0.3413

From the tabulations we are able to note that, there is a non significant effect of age both within the sections and across the MIS sections:

Figure 6: A plot of the estimated variance against age

A closer look at the variance components though non significant we are able to reveal that; individuals below the mean age have a high variation in gross family income compared to those

above the mean age. In fact the lower the age an individual has the more variant is the gross family income across the sections as it can be depicted by figure 6 below.

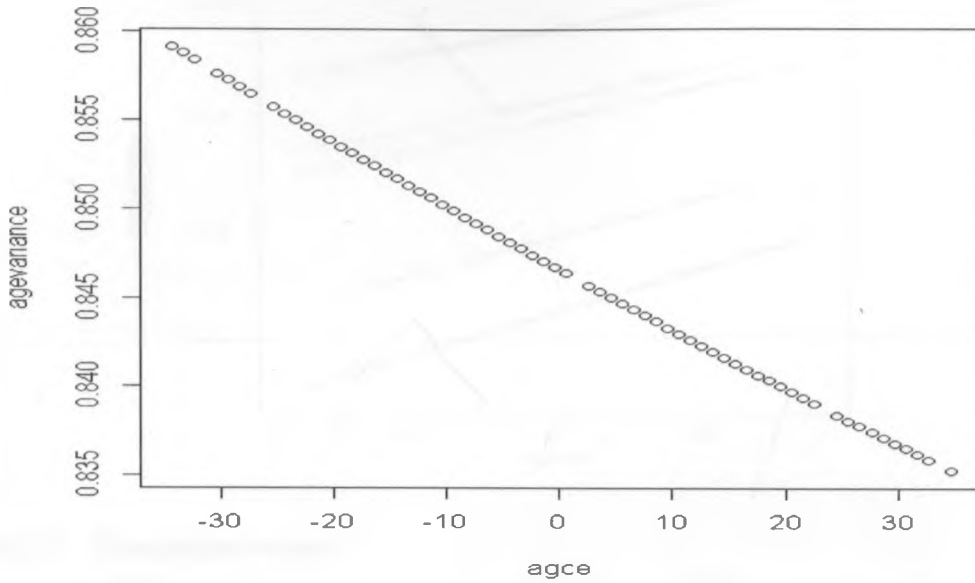
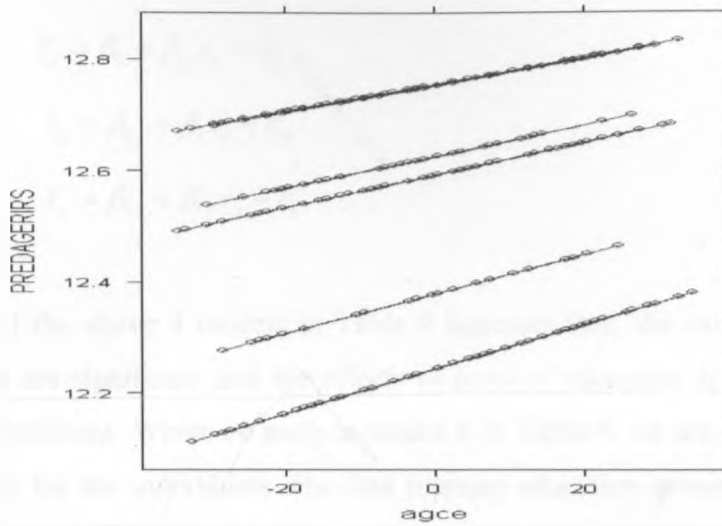


Figure 7: A plot of estimated income against age

The figure 7 below (page 44) shows the regression lines for the sections when the slope effect is assumed to be the same for all the sections (model 3 in Table 7). The figure shows that there is direct relationship between age of persons in these sections and the level of Gross family income such that the higher the age of the person the higher is the level of Gross family income and vice versa. Sectional intercept is significant implying there is a difference in Gross family income across the sections, but the slope effect is found to be non significant in this case implying the difference in gross family income between those who are perceived to be old and the young ones is not significant.

figure7



4.2.1.2.3 Education effects

Table 9: Multilevel models with level of education as the explanatory variable

	Null model or Empty model	Fixed intercept Random slope model	Random intercept fixed slope model	Random intercept random slope model
	Model 1	Model 2	Model 3	Model 4
Fixed part				
intercept	12.56 (0.11)	12.544 (0.079)	12.511 (0.12)	12.526 (0.161)
level of education	—		0.191(0.117)	0.01957 (0.191)
Random part				
τ_{00}^2	0.052703	0.11619	0.064914	0.13471
τ_{10}				-0.1393576
τ_{11}^2		0.11894		0.14416
σ^2	0.791860	0.75252	0.785669	0.75324
ICC	0.0624027			

Equations for the models in Table 9

Model 1	$Y_{ij} = \beta_{0j} + \varepsilon_{ij}$	
Model 2	$Y_{ij} = \beta_0 + \beta_{1j}x_{ij} + \varepsilon_{ij}$	
Model 3	$Y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij}$	4.9
Model 4	$Y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij}$	

A summary of the above 4 models in Table 9 indicates that, the variance components for the above models are significant and the effects of level of education is non significant while the intercept is significant. When we analyze model 4 in Table 9, we are able to note that the gross family income for the individuals who had primary education across the MIS sections is $\exp(12.526) = \text{Kshs.}275405.5$ and for those individuals with education level beyond primary had gross family income mean of $\exp(12.526+0.01957) = \text{Kshs.} 280848.3$, hence the mean ‘gross family income differential’ is $\text{Kshs.} 5442.8$. The gross family income differential is no longer constant across sections, but varies by the amount u_{1j} around the mean, γ_{10} and this figure is non significant.

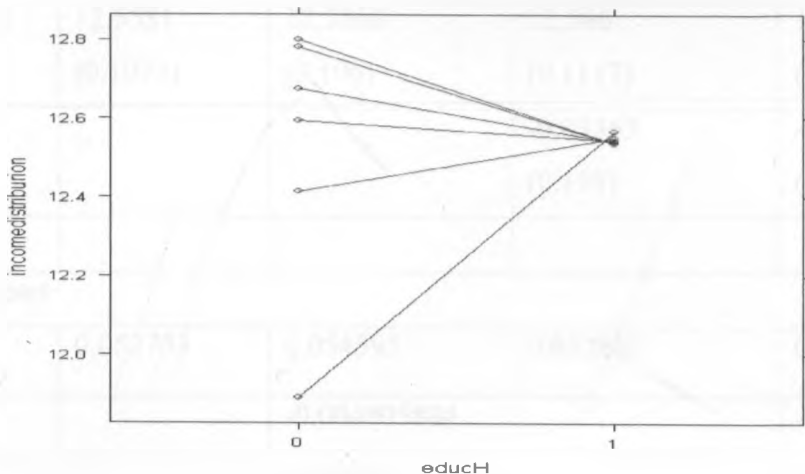
The variation in gross family income across the sections after accounting for the level of education is given by;

$$\text{vay}(\log \text{income})_{ij} = \tau_{00}^2 + \text{educ}_{ij}^2 \tau_{11}^2 + 2\text{educ}_{ij} \tau_{10} + \sigma^2 \quad 4.10$$

The variance in gross family income for those individuals with primary level of education within the sections is $(0.13471+0.75324) = 0.88795$ unlike for the individuals with past primary education the variance in gross family income within the sections is $(0.13471+0.14416+ (-0.1393576*2) +0.75324) =0.7533948$. The marginal variation between the two groups is non significant as can be revealed by testing model 3 against model 4 in table 9.

Figure 8: A plot of the estimated income against level of education

When we fit the independent random intercept random slope model where we don't let the intercept and the slope to co vary, we find that there is a high variation of Gross family income for those individuals with primary level of education at 0.13471129 as compared to the individuals with past primary education which is at 0.00016005 as depicted by figure 8 below.



4.2.1.2.4 Effects of marital status

Table 10: Multilevel models with marital status as the explanatory variable

Model	Null model Or Empty model	Fixed intercept Random slope model	Random intercept fixed slope model	Random intercept random slope model
	Model 1	Model 2	Model 3	Model 4
Fixed parts				
Intercept	12.5581 (0.1072)	12.5568 (0.106)	12.566 (0.1117)	12.56797 (0.114)
marital status	–		-0.03363 (0.119)	-0.03565 (0.119)
Random part				
τ_{00}^2	0.052703	0.054595	0.053283	0.056015
τ_{10}		-0.004905804		-0.006883961
τ_{11}^2		0.000441		0.00084601
σ^2	0.791860	0.791862	0.794138	0.794123
ICC	0.0624027			

Model specifications in Table 10

Model 1 $Y_{ij} = \beta_{0j} + \varepsilon_{ij}$

Model 2 $Y_{ij} = \beta_0 + \beta_{1j}x_{ij} + \varepsilon_{ij}$

Model 3 $Y_{ij} = \beta_{0j} + \beta_1x_{ij} + \varepsilon_{ij}$

Model 4 $Y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij}$

4.11

An investigation of the models in Table 10 reveals that there is non significant difference in gross family income between individuals who are married and those not married in all the

models i.e. (marital effect non significant). But there is a significant difference in gross family income between the sections as can be shown by a significant effect of the intercept.

A closer look at the variance components indicates that though the marital effect is non-significant, there is some degree of variation in gross family income among the individuals and across the sections in the random slopes models. The variation in gross family income for those individuals that are married is higher at 0.8911358 as compared to those not married 0.837216 within the sections and similarly when we fit an independent random intercept random slope model we get to see that the same kind of situation prevails at the section level where the sectional variance for the married is 0.056013 as compared to the not married individuals at 0.043085 gross variation. The variation difference is non significant.

4.2.1.2.5 Effects of the demographic factors

Table 11: Fixed part of Multilevel Models with Demographic factors as the Explanatory Variable

Model	Null model Or Empty model	Fixed intercept Random slope model	Random intercept fixed slope model	Random intercept random slope model
	Model 1	Model 2	Model 3	Model 4
Fixed parts				
Intercept	12.56 (0.1)	12.5410(0.069)	12.514734(0.128)	12.5552(0.1841)
Gender	—		-0.0621(0.128)	-0.0315(0.265)
Centred age			0.002719(0.0031)	0.003(0.003)
Marital status			0.04927(0.173)	-0.0111(0.204)
Level of education			0.1825(0.1192)	0.0006(0.2035)

The Models in Table 11 are specified as follows;

Model 1
$$Y_{ij} = \beta_{0j} + \varepsilon_{ij}$$

Model 2
$$Y_{ij} = \beta_0 + \sum_{k=1}^4 \sum_{i=1}^{n_j} \beta_{kj} x_{kij} + \varepsilon_{ij}$$

Model 3
$$Y_{ij} = \beta_{0j} + \sum_{k=1}^4 \sum_{i=1}^{n_j} \beta_{kj} x_{kij} + \varepsilon_{ij}$$

Model 4
$$Y_{ij} = \beta_{0j} + \sum_{k=1}^4 \sum_{i=1}^{n_j} \beta_{kj} x_{kij} + \varepsilon_{ij}$$

4.12

When we look at the fixed parts of the models in Table 11, we can note that the intercept is highly significant and the effects of all the explanatory variables that constitute demographical factors are non significant both within the sections and across the sections.

4.2.1.3 Modeling institutional factors

4.2.1.3.1 Effects of extension services

Table 12: Multilevel Models with Extension Services as the Explanatory Variable

Model	Null model or Empty model	Fixed intercept Random slope model	Random intercept fixed slope model	Random intercept random slope model
	Model 1	Model 2	Model 3	Model 4
Fixed parts				
intercept	12.5581(0.11)	12.5499(0.11)	12.60443(0.128)	12.60(0.1442)
extension services	-	-	-0.06978(0.11)	-0.0666(0.126)
Random part				
τ_{00}^2	0.052703	0.6989	0.051623	0.07662
τ_{10}		-0.01851189		-0.02700168
τ_{11}^2		0.01483		0.022285
σ^2	0.791860	0.78879	0.793602	0.78967
ICC	0.0624027			

Model specifications for the above Table 12

Model 1 $Y_{ij} = \beta_{0j} + \varepsilon_{ij}$

Model 2 $Y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}$

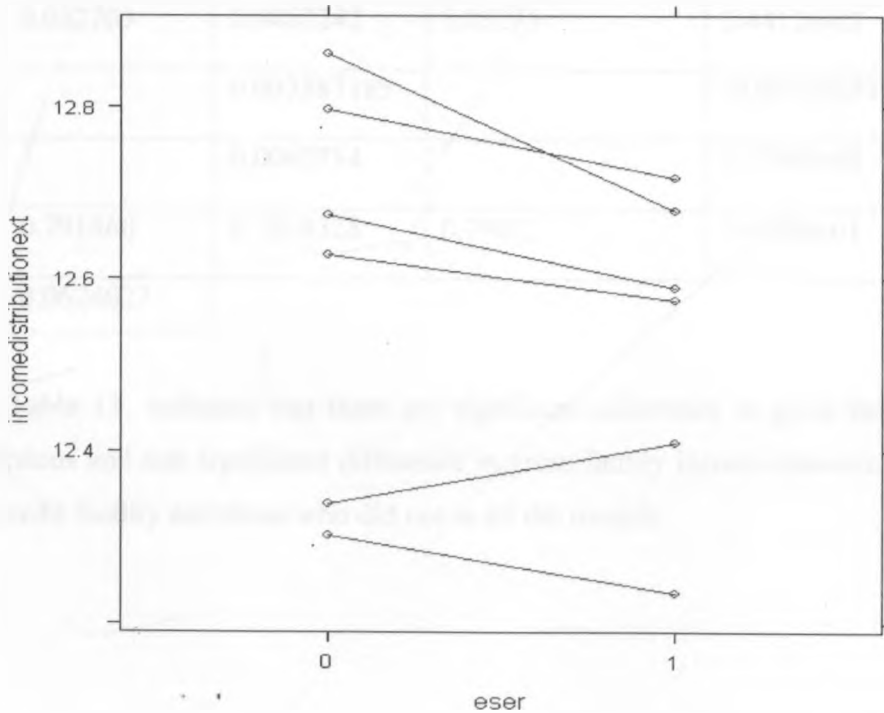
4.13

Model 3 $Y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \varepsilon_{ij}$

Model 4 $Y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \varepsilon_{ij}$

Figure 9: A plot of estimated gross family income against extension services

The models in Table 12 indicate a significant intercept and a non significant effect in extension services. This suggests that, there are significant differences in gross family income between the sections but there is non significant difference in gross family income between individuals who sort help from agricultural officers and those who did not. Although this is the case, we are able to note that, individuals who sort help from extension officers had slightly higher gross family income across the sections except for Karama section as depicted in figure 9 below.



4.2.1.3.2 Effects of access to credit

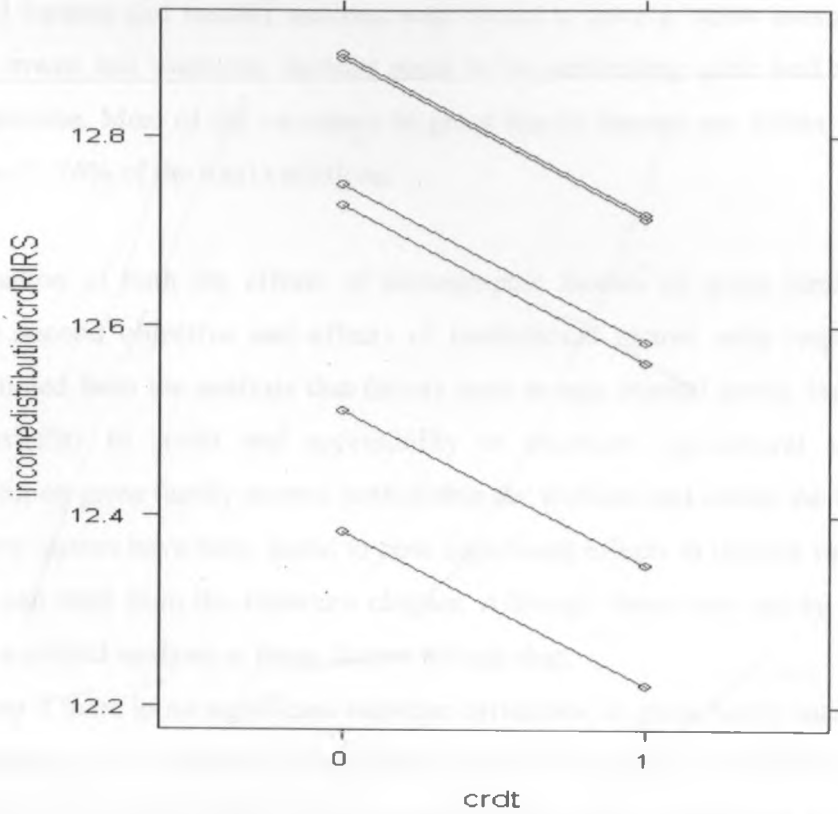
Table 13: Multilevel Models with access to Credit as the Explanatory Variable

Model	Null model Or Empty model	Fixed intercept Random slope model	Random intercept fixed slope model	Random intercept random slope model
	Model 1	Model 2	Model 3	Model 4
Fixed parts				
intercept	12.5581(0.1072)	12.564(0.107)	12.6895(0.1484)	12.6898(0.1494)
Access to credit	-	-	- 0.1652(0.1292)	-0.1657(0.1292)
Random part				
τ_{00}^2	0.052703	0.0467242	0.05275	5.4412e-02
τ_{10}		0.003587185		-0.001036571
τ_{11}^2		0.0002754		1.9747e-05
σ^2	0.791860	0.7918328	0.79021	7.9020e-01
ICC	0.0624027			

The results of Table 13 indicates that there are significant difference in gross family income between the sections and non significant difference in gross family income between the persons who accessed credit facility and those who did not in all the models.

Figure 10: A plot of estimated income against access credit

Figure 10 show the slope is negative for the regression lines of each of the sections. The persons who accessed credit facilities had a higher gross family income as compared to the persons who did not access this facility, but the difference between them is non significant as can be revealed the by the insignificant regression coefficient of effect to access to credit as shown in Table 13 .



CHAPTER FIVE:

CONCLUSION AND DISCUSSIONS

The analysis stage has shown there is a significant different in gross family income distribution across the sections and looking at our first objective we are able to say that there exists a difference in the gross family income across the sections of MIS. A test of sectional effects was significant where we saw that 6.24% of the variations in gross family income are due to sectional differences. The variation of 0.845 in gross family income across the sections was found to be significant and karama and mutithi sections were found to have a below average gross family income while mwea and wamumu sections seem to be performing quite well with the highest gross family income. Most of the variations in gross family income are within the sections and this constitutes 93.76% of the total variations.

In the examination of both the effects of demographic factors on gross family income with respect to the second objective and effects of institutional factors with respect to the third objective, we noted from the analysis that factors such as age, marital status, level of education, gender, accessibility to credit and accessibility to extension agricultural services had no significant effect on gross family income both within the sections and across the MIS sections.

All of the above factors have been found to pose significant effects in income variations in other studies as we can read from the literature chapter. Although these may not be the case for the current study, a critical analysis at these factors reveals that;

- i. Even if there in no significant variation difference in gross family income both within and across the sections for individuals with different levels of education, we get to see that for those individuals who proceeded past primary level had a lower gross family income variance of 0.7533948 as compared to 0.88795 for those who had primary education or did not have education at all within the sections and across the sections they had 0.13471129 and 0.00016005 respectively.
- ii. There was no significant difference in gross family income for the individuals below the average age and those above this mark. A look at the variations in gross family income reveals that there is an inverse relationship between age and variation in gross family income across the sections such that variation in gross family income decreases with an increase in age as depicted by figure 7. Hence, we can say that

gross family income is directly inversely proportional to age across the sections, although this variation is statistically non significant.

- iii. There is no significant difference in gross family income between married and not married individuals both within and across the sections. There is greater variability in gross family income for those individuals who are married both within and across the sections as compared to those not married although the variation is not significant.
- iv. When we look at accessibility to agricultural extension services, we note that there is no significant difference in gross family income between those individuals who accessed help of agricultural officers and those who did not both within and across the sections although for those who sought help had a marginally higher gross family income. When we assess the variations, we find that those individuals who sought guidance had a greater variation in gross family income of 0.86629 compared to those who did not of 0.8345716 within the sections. Across the sections the variation showed the same trend where for those who sought guidance the variation was 0.0342047 compared to the others at 0.0030526.
- v. Access to credit have been found to be of significant effect in studies related to agricultural productivities such as indicated in the literature section, but for this study it has been found to be non significant. The difference in gross family income between those individuals who accessed credit and those who did not is insignificant; even though those who did access credit had a marginally higher gross family income compared those who did not. When we look at the variance of Gross family income within the sections for each of this two groups, we get to see that it's equal at 0.844612 but the across sections variance for the two groups is zero.
- vi. The effect of gender is non significant in all the types of the model, but we are able to notice that the variation in Gross family income is higher for males than for females headed house holds across the sections. When we look at the sections variations in gross family income by section, we see that in karama and mutithi sections female headed house holds have a higher Gross family income than for households headed by males. For the other sections the females headed households have a lower gross family income.

When we plot an independent random intercept random slope, we are able to note while some sections exhibit a negative gradient for the slope with respect to the factor effect under consideration, there are some sections that have a positive gradient. For example;

- i. The males from karama and mutithi sections have a slightly lower gross family income as compared to their female counterparts unlike the other sections where males have a higher gross family income.
- ii. Individuals in Karama seem not to impress the extension services or a lack of know how on the importance of these services while the other sections impress these services especially Wamumu section.
- iii. Education seems to be a serious issue in all the sections with less than 30% having gone over the primary level. This poses a big challenge in the long term production of rice, necessitating effort by the government to do expansion and the capacity building programs offered by NIB. Although only in two sections Mutithi and Karama has education shown a slight positive significant effect, the other sections show the opposite and the good thing is for those individuals with past primary education they had a slightly converged Gross family income across the sections unlike the other group.
- iv. Across the sections we can note that from figure 12 those individuals who accessed credit facilities had a slightly higher Gross family income as compared to those who did not although the difference was insignificant.

When we look at the standard errors of the u_{0j} 's or sectional residuals, we note that these estimates are not really significant and this problem may have a significant effect on estimations of other parameters in the analysis and hence in our results interpretations.

REFERENCES

1. Snijders, T. A. & Boskers, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks: Sage Publications
2. Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*, 2nd Edition. Thousand Oaks, CA: Sage.
3. Pinheiro, J. C. & Bates, D. M. (2000). *Mixed effects models in S and S-Plus*. New York: Springer-Verlag.
4. Aitkin, M. & Longford N. (1986). Statistical modeling in school effectiveness studies. *Journal of the Royal Statistical Society, A*, 149, 1–43.
5. Jerry W. D. and Jeffrey R. W. (2000), farm characteristics that influence net farm income variability and losses, Department of Agricultural Economics, Kansas State University, Manhattan.
6. Laird N. M. & Ware J. J. (1982), Random-effect models for longitudinal data. *Biometrics*, 38, 963-974.
7. Central Bureau of Statistics (2010), Kenya Demographic and Health Survey Report.
8. Saito K. A., Mekonnen H. & Spurling D. (1994), Raising productivity of women farmers in sub-Saharan Africa. World Bank Discussion Paper 230. Washington, DC.
9. Maitha J. K. & Senga W.M. (1976), *Agricultural development in Kenya; economic assessment*, Oxford presses.
10. Kuria L.M. (2001), Factors hindering the fostering of entrepreneurship in agriculture : a case study of Mwea Irrigation Scheme rice farmers, Jomo Kenyatta University of Agriculture And Technology, Kenya.
11. Achia T. N., Wangombe A.& Khadioli N. (2010), A Logistic Regression Model to Identify Key Determinants of Poverty Using Demographic and Health Survey Data, *European Journal of Social Sciences – Volume 13, Number 1*
12. Goldstein H. (1995), *Multilevel Statistical Model*
13. Tavneet S., Tschirley D., Irungu C., Gitau R. & Kariuki D. (2008), *Rural Incomes, Inequality And Poverty Dynamics In Kenya*
14. Kuria J. N., Ommeha H., Kabuageb L., Mbogoc S. & Mutero C. (2008), *Technical Efficiency of rice producers in Mwea Irrigation Scheme*

15. Mathenge M. & Tschirley D.L. (2008), Income Growth and Mobility of Rural Households in Kenya: Role of education and historical patterns in poverty reduction. Paper presented at CSAE 2008 Conference on Economic Development in Africa, St. Catherines College, Oxford, 16-18 March 2008.
16. Bock (1989), *Multilevel Analysis Of Educational Data*, San Diego, Academic Press.
17. Quisumbing R. (1994) *Gender Difference In Agricultural Productivity: A Survey Of Empirical Evidence, Education And Social Policy Discussion Paper No. 36*, World Bank Washington D.C
18. Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models for social and behavioural research: Applications and data analysis methods*. Newbury Park, CA: Sage Publications
19. Holt, Scott & Ewing (1980). *Hierarchical Linear Model*. Sage publications
20. Jennrich, R. & Schluchter, M. (1986). Unbalanced repeated measures models with structured covariance matrices, *Biometrics* 42: 805–820.
21. Gibbons R. D. & Bock R. D. (1987), *Trend in correlated proportions*
22. Gitau M. J. (2005), *How to understand inequality in Kenya*, Partner News
23. Bliese P. (2012), *Hierarchical Linear Modeling*