بسم الله الرحمن الرحيم

Sudan University of Science and Technology

College of Graduate Studies

Use of Multiple Logistic Regression to Estimate the Effect of Socio-Economic Factors on Household Income Sufficiency

(Case study: South Darfur state)

إستخدام الإنحدار اللوجستي المتعدد لتقدير أثر العوامل الاقتصادية والإجتماعية على كفاية دخل الأسرة

(دراسة حالة ولاية جنوب دارفور)

A thesis submitted in fulfilment of the

requirements for the ph.d in statistics

Prepared by:

Sofian Abuelbacher Adam Saad

Supervisor:

Dr. Amin Ibrahim Adam

June. 2016

الآية

بسم الله الرحمن الرحيم

قال تعالى جل في علاه:

وَقُلِ اعْمَلُوا فَسَيَرَى اللَّهُ عَمَلَكُمْ وَرَسُولُهُ وَالْمُؤْمِنُونَ وَسَتُرَدُّونَ إِلَى عَالِمِ الْغَيْبِ وَالشَّهَادَةِ فَيُنَبِّئُكُمْ بِمَا كُنْتُمْ تَعْمَلُونَ "

صدق الله العظيم

سورة التوبة {105/9}

Dedication

To my kids

Alaa, Ahmed and Mohamed

Acknowledgement

I gratefully acknowledge my honest supervisors the A. Professor. Amin Ibrahim Adam, the dean faculty of economics, Omdurman Islamic University, and the Dr. Afraa Hashim Abdelateef, vice dean faculty of Sciences, Sudan University of Sciences and Technology for their fruitful support to come out by this study. I owe special thanks to all people those who are kindly encourage and provide me by their valuable comments and advices, in particular the mother, father, wife and my colleagues at the University of Nyala.

Abstract

The main objective of this study is to find the main factors that affect household income sufficiency in Sudan, since there were many economic variants that happened during the past ten years representing in the secession of the south part of the country which leads the country to lose more than 70% of its oil production and seriously affect the level of the national income.

The data are collected from south Darfur state in 2015 using cluster random sample techniques covering twenty one localities'.

The problem of the study represented on how we can come out by the main factors that affected the sufficiency of household's income in Sudan taking South Darfur state as a case of study during the period of time 2013 to 2016.

To achieve the objective of the study, nine variables (Household size, household income level, household expenditure level, household head age, household head education level, household head gender, household head type of occupation, students at school, and student at the university) have been used in order to identify which of which has the most impact on the sufficiency of household income.

The method of multiple logistic regression is used and the data are analyzed by the tool of statistical package (SPSS) after fulfillment of the requirements of the methodology in respect to the type of data.

The results of the study indicated that there are five variables that have a significant impact on the dependent variable (household income sufficiency). These are; the educational level of the household head, the household size, the household level of expenditure, the level of monthly

income and the number of students at the university level. A household to be within the less satisfy income group has a largest probability (>= 0.766) in comparing to be within the other groups (i.e. to some extent satisfy income group and the quite satisfy income group). The main recommendations of this study is to raise the level of education of the household's heads to improve the economic situation for their families and encourage them to take care of their family size.

مستخلص البحث

هدفت هذه الدراسة الى تحديد أهم العوامل والمتغيرات التي تؤثر على كفاية دخل الاسرة فيما يلي توفير المتطلبات الاساسية للعيش الكريم بالمجتمع السوداني خاصة في ظل عدد من المتغيرات الاقتصادية المؤثرة التي طرأت على البلاد خلال العشرة اعوام الماضية و التي على رأسها إنفصال دولة جنوب السودان عن شماله الشئ الذي أدي فقدان الدولة لحولي اكثر من 70% من منتجات البترول مما ادى الي تدهور حاد في مستوي الدخل القومي الذي اثر بدوره علي كفاية دخل الاسر.

تم جمع البيانات من ولاية جنوب دارفور في العام 2015م بإستخدام أسلوب العينة العنقودية لتغطية إحدى و عشرون محلية بالولاية.

و قد تمثلت مشكلة الدراسة في كيفية التعرف على أهم العوامل التي تؤثر على كفاية دخل الاسرة بالتركيز على ولاية جنوب دارفور كدراسة حالة خلال الفترة من عام 2013 الي 2016م،

لتحقيق الهدف من الدراسة، إستخدمت الدراسة تسعة من المتغيرات (مستوي دخل الاسرة، مستوى الإنفاق، حجم الأسرة، عمر رب الأسرة، المستوى التعليمي لرب الأسرة، عدد الأبناء بالمدارس، الأبناء بالجامعات، نوع مهنة رب الأسرة و جنس رب الأسرة) بغرض التعرف على أكثرها تأثيرا على كفاية دخل الاسرة.

للوصول للنتائج المرجوة تم إستخدام منهجية الانحدار اللوجستي المتعدد بالإستعانة بالحزمة الإحصائية للعلوم الإجتاعية (SPSS) وذلك بعد استيفاء كافة متطلبات المنهجية فيما يتعلق بنوع البيانات المستخدمة.

أظهرت نتائج التحليل أن هناك خمسة عوامل كانت ذات أثر جوهري على كفاية دخل الأسرة وهي المستوى التعليمي لرب الأسرة، حجم الأسرة، مستوى الإنفاق الشهري، مستوى الدخل الشهري و عدد الأبناء بالجامعات.

أظهرت نتائج التحليل ايضا أن هناك احتمالا مقداره 0.766 أو أكثر على أن أي أسرة يتم اختيارها من مجتمع الدراسة يمكن ان تقع ضمن مجموعة الأسر ذات الدخل الغير كافي مقارنة بوقوعها في المجموعتين الأخريين (مجموعة الأسر ذات الدخل الكافي و مجموعة الأسر ذات الدخل المتوسط).

كما خرجت الدراسة بعدد من التوصيات أهمها رفع المستوى التعليمي لأرباب الأسر وذلك بغرض تحسين الأوضاع الإقتصادية لأسرهم ,و تشجيعهم على ضرورة الإهتما بالحفاظ على حجم أسرهم بما يتماشى مع أوضاعهم الإقتصادية.

Table of contents

Table of tables

Table of figures

Chapter One

Introduction

1-1 Preface

1-2 Problem of the study

1-3 Significance of the study

1-4 Objectives of the study

1-5 Hypothesis of the study

1-6 Methodology

1-7 Limits of the study

1-8 previous studies

1-9 Structure of the study

1.1 Preface

Sudan is an extremely poor country that has experienced protracted social conflict, civil war, and, in July 2011, the loss of three-quarters of its oil production due to the secession of South Sudan. The oil sector had driven much of Sudan's GDP growth since 1999. For nearly a decade, the economy boomed on the back of rising oil production, high oil prices, and significant inflows of foreign direct investment. Since the economic shock of South Sudan's secession, Sudan has struggled to stabilize its economy and make up for the loss of foreign exchange earnings. The interruption of oil production in South Sudan in 2012 for over a year and the consequent loss of oil transit fees further exacerbated the fragile state of Sudan’s economy. Sudan is also subject to comprehensive US sanctions. Sudan is attempting to develop non-oil sources of revenues, such as gold mining, while carrying out an austerity program to reduce expenditures. The world’s largest exporter of gum Arabic, Sudan produces 75-80% of the world’s total output. Agriculture continues to employ 80% of the work force. Sudan introduced a new currency, still called the Sudanese pound, following South Sudan's secession, but the value of the currency has fallen since its introduction. Khartoum formally devalued the currency in June 2012, when it passed austerity measures that included gradually repealing fuel subsidies. Sudan also faces rising inflation, which reached 47% on an annual basis in November 2012, but subsided to 25% in 2013. Ongoing conflicts in Southern Kordofan, Darfur, and the Blue Nile states, lack of basic infrastructure in large areas, and reliance by much of the population on subsistence agriculture keep close to half of the population at or below the poverty line. However, all these things lead to sever increase in commodities prices in

the relevant period of time which highly affected household ability to meet their basic needs; this is together with the limitation of their sources of income.

Income is by no means the only way to support consumption and/or other types of expenditure, as financial assets can be run down and real assets can also be used to generate liquidity (reverse mortgages, equity lines etc.).

However, social scientist and economists have always shown a keen interest in income, for instance in their studies of economic inequality and poverty , and in most health surveys containing questions on economic and social wellbeing, the only measure of access to economic resources is income. Indeed, income is an important (arguably, the most important) component of any measure of access to economic resources, thus deserving careful investigation on it is own.

In broad terms, income refers to receipts, whether monetary or in kind, those are received at annual or more frequent intervals and are available for current consumption. For most people, household income is the most important determinant of economic well-being. Household income provides a measure of the resources available to the household for consumption and saving.

Scientists provided more many definitions concerning the related concepts of households and income; Household income is total income from all people living in a particular household. Income refers not only to the salaries and benefits received but also to receipts from any personal business, investments, dividends and other income. Furthermore, household members do not need to be related to be part of a household. Household income is often used as an economic indicator.

However, the total Income is the sum of all money received by an individual or organization, including income from employment or providing services, revenue from sales, payments from pension plans, income from dividends, or other sources. Total income may be calculated for purposes of assessing taxes, evaluating the net worth of a company, or determining an individual or organization's ability to make payments on a debt.

Household represent all persons living under one roof or occupying a separate housing unit, having either direct access to the outside (or to a public area) or a separate cooking facility. Where the members of a household are related by blood or law, they constitute a family.

On the disbursements side of household accounts, consumption expenditure represents the day-to-day purchases that may be financed not only by household income but also by savings from previous periods or by incurring debt. For some households, such as retired households, the running down of capital for consumption may represent a deliberate attempt on their part to even out consumption over a lifetime. Other groups in the population, such as farmers, may also average out their consumption over a number of years, while their incomes may show quite wide fluctuations over the same period. In such cases, consumption expenditure may represent a better estimate of the household’s sustainable standard of living (better measure of income instead).

Household income was found to be an effective tool that measure families' economic well-being.

Income sufficiency considered as a main factor that affects household expenditure, at the same time insufficiency of the income would actually negatively affected both the quality and quantities of goods that

demanded by the families. In addition to that the prices would be another important factor determines the household consumption.

Generally household income play a fundamental role in the economic development and help countries decision makers to draw right development plans in one hand, and on the other hand it might help related development actors to fairly distribute the general income and government subsidies among the whole society.

Although there are more than study discussed the general economical and social variables of the household, but studies regarding household income are very rare, this situation encourage undertaking this study to determine the Sudanese household income level particularly when there are a wide range of goods prices instability throughout the previous three decades of time.

1.2 Problem of the study

The problem of the study concentrate mainly on how can we come out by the main effective variables that affected the efficiency of the households income and to generate statistical model showing the relationships between the response variable and the explanatory variables in terms of the odds ratio.

1.3 Importance of the Study

- To find out all the socio-economic factors that affected the insufficiency of the household income
- To find out suitable alternative economic means and resources that would lead to increment of the household income especially after the country loses of more than 70% of the oil as one of the most important economic resources that Sudanese societies was relied

on to increase and supplement the family income, after the secession of southern Sudan recently in the mid of 2011.

- Almost all Sudanese society (urban & rural) would benefit more from this study
- It can be very easy to obtain the required data for the study, in terms of availability, means of collection, types of the data and equipments… etc.
- The research can be completed in the proposed time period. (i.e. PhD research)

1.4 Objectives of the study

The most important characteristic of this study is to focus on the uses of multinomial logistic regression on some economic applications since the multinomial logistic regression analysis is commonly used in the fields of medical and social studies where much of the data in such fields are often in the shape of binary response.

1. To study and discuss the importance of the application of multinomial logistic regression in the analysis, where the nature of the elected data was categorical.
2. To estimate the determinants that would lead to insufficiency of household income.
3. To create Statistical Model that will help determine the level of households' welfare.
4. Demonstrate the importance and utility of multinomial logistic regression.
5. Assisting in drawing the high level state's policies and strategies towards social and economic issues

6. To identify the most important reasons that lead to insufficiency of household income.
7. Analyzes the strength of location determinants of rural community income, in order to ascertain what may have caused differences in income levels among the surveyed community.
8. Interpret the results of using multinomial logistic regression model in the analysis of categorical data based on the concepts of the odds ratio.

1.5 Hypotheses of the study

The research study will adopt the following hypotheses:-

1. There is no significance effect of the household head age on the dependent variable**.**
2. There is no significance effect of the household size on the dependent variable)**.**
3. There is no significance effect of the household level of expenditure on the dependent variable)**.**
4. There is no significance effect of the household's number of student at university level on the dependent variable)**.**
5. There is no significance effect of the household's number of student at school level on the dependent variable)**.**
6. There is no significance effect of the household monthly of income on the dependent variable)**.**
7. There is no significance effect of the household head year of education on the dependent variable)**.**
8. There is no significance effect of the household head type of occupation on the dependent variable)**.**

9. There is no significance effect of the household head gender on the dependent variable)**.**

1.6 Methodology of the study

The study focused on Multinomial Logistic Regression (MLR) as an inferential statistical method to identify the relationship between a categorical response (dependent) variable having more than two levels of categories and the explanatory (independent) variables, whether these variables are numeric or categorical or both.

The goal of multinomial logistic regression is to construct a model that explains the relationship between the explanatory variables (i.e. household size, educational level, occupation, age, gender and type of living house) variables and the outcome (level of income).

1.6.1 Sampling and data collecting

The sample framework of the study shall be the rural Societies where most of the Sudanese rural inhabitants were living below the poverty line with significant insufficient sources of income. In particular the zone areas of the study will take place at the South Darfur state.

South Darfur state has twenty one localities. According to the purpose of the study the localities will be divided in to three clusters, seven localities in each cluster, two clusters will be selected at random. For the sake of good sample representation and high data quality, the sampling frame inside each locality is selected to be stratified sampling techniques since the number of villages in each localities not equally likely, this is in order to verify which villages to be among the sample in order to cover the distribution of both geographical dispersion and households' variability. For selection of household, also the sampling frame inside

each villages is selected to be stratified. From each village a number of households will be selected at random to form a total sample of 307 households, which seems to be adequate to pursue the objectives of the study. The number of villages in each locality is linked to the number of households, which is determined in proportion to the locality population according to the Sudan Population Census of 2008.

1.7 Limits of the study

1. Location of the study is South Darfur State
2. Time framework of the study is during the period: 2013 - 2016

1.8 Previous studies

During the past ten years, many studies had been used the Multinomial Logistic Regression (MLR) to analyze categorical data. These studies had varied on different subjects including those related to social and medical issues, behavioral, and some scientific experiments. The purpose of reviewing these studies was to explore the practical, technical methods, and statistical treatments had been used. So the focus was on these methods in practice that used the logistic regression model.

George Udny Yule (1871-1951), presented the odds ratio and related measures of association, before then, most work focused on descriptive aspects for relatively simple measure.

In 1933 report on quantal response methods by J. H. Gaddum, popularized the probit model for applications in toxicology with a binary response.

Chester Bliss (1935), introduced the term probit but he used the inverse of normal cumulative distribution function (cdf) with mean 5, rather than 0, in order to avoid negative values finding and standard deviation 1.

In the same year, 1935, Fisher outlined an algorithm for maximum likelihood (ML) estimates of model parameters.

Bartlett (1935) showed how to find ML estimates of cell probabilities satisfying the property of equality of odds ratios between two variables at each level, he attributed the idea to Fisher.

Bartlett (1937), used $\log = \left[\frac{y}{1-y} \right]$ in regression and analysis of variance (ANOVA) to transform observations y that are continuous proportions.

Joseph Berkson (1944), introduced the term logit for this transformation, Berkson showed that the model using the logit fitted similarly to the probit model, and his subsequent work did much to popularize logistic regression.

In 1951, Jerome Cornfield, with strong medical ties, used the odds ratio to approximate relative risks in case control studies.

After one year, in 1952, Dyke and Patterson, apparently first used the logit in models with qualitative predictors.

David R. Cox introduced logistic regression, through his 1958 article and 1970 book" The Analysis of Binary Data".

At the same time, an article by George Rasch (1958) sparked an enormous literature on item response models; the most important of these is the logit model with subject and item parameters, now called the Rasch model. This work was highly influential in the psychometric community of northern Europe and spurred many generalizations in the educational testing community in the United States.

Nathan Mantel (1959) made a variety of interesting contributions to Categorical data Analysis (CDA). Nathan Mantel also discussed trend tests, and in 1966, he discussed multinomial logit and loglinear modeling, also, in 1973, he discussed logistic regression for case control data.

Chao and Rebecca (2002), demonstrated the utility of multinomial logistic regression (MLR) model to identify adolescent at greatest health risk from their personal as well as family characteristics. They used the model to predict the likelihood of a categorical response variable, using a sample of 432 students enrolled in two junior high school (grades 7 through 9). The response variable was students' risk level on the behavioral risk scale with three levels (high risk, medium risk, and low risk). Explanatory variables included gender (two categories), intention to drop out of school (two categories), and family structure (with three categories). The research hypothesis posed to the data was stated as follows: the likelihood that an adolescent is at high, medium, or low behavioral risk is related to his/her gender, intention to drop out of school, family structure, emotional risk and self-esteem. The study used Statistical Analysis System software (SAS) to calculate MLR. Model was validated by significant test of overall model and tests of regression parameters, goodness-of-fit measures and validation of predicted probabilities. The study had been divided into four sections: description of the data and research question, MLR, interpreting and assessing of MLR, and finally, summary.

Nichols and et al, (2005), in a study compared self-reported dry eye disease across contact lens wearers, spectacles wearers, and clinical emmetropes (those not requiring refractive correction). Response variable was mode of refractive correction with three categories: contact lens wearers, spectacle wearers, and clinical emmetropes. Sample size was

(n=893). The explanatory variables were: gender, age, dryness, light sensitivity, and self–reported. Researchers used special scoring to build the dryness and light sensitivity variables. Chi-squared test was used to determine the relation between gender and mode of correction. Logistic regression was used to examine the relation between mode of refractive correction and self-reported dry eye disease, using the cutoff scoring algorithm, controlling for age and gender. Adjusted odds ratio (OR), 95% confidence interval, and probability were reported. Hosmer-Lemeshow goodness-of-fit test was used to examine the final calibration of the model. When the chi-squared value for this test is small (high probability), the model is considered well calibrated, the discriminative ability of the models was evaluated using the area under the receiver operating characteristic (ROC) curve, discrimination was assessed using the following guidelines for area under the ROC curves: 0.5 indicated no discrimination, between 0.7 and 0.8 indicated acceptable discrimination, between 0.8 and 0.9 indicated excellent discrimination, and greater than 0.9 indicated outstanding discrimination. A multinomial logistic regression was then used to model the relation between the frequency of each symptom and mode of refractive correction.

Woo and Ditton (2006), in certain study they explored the relationships among variables regarding the willingness to substitute one location for another location. They determined four types of substitution, alternatives, involving resource and activity alternatives (this was the response variable). In the case of recreational fishing, anglers may choose one of these four types of substitution alternatives to get the same recreational satisfaction and benefits they got from fishing. Most importantly, this study explored the relationship between specialization variables (behavior, skill/ knowledge, and commitment) and anglers' willingness to

substitute. The response variable had four categories and the explanatory variables were 16 variables included recreation specialization, demographic, and constraints variables, from these explanatory variables, there were three numeric variables. The sample size was 1005 observations (n=1005). Results of MLR were showing significant effects on anglers' willingness to substitute other locations (final model included only significant variables at .05 levels). MLR provided sufficient evidence that recreation specialization was closely associated with an individual's willingness to substitute other locations as in a previous study, which means that recreation specialization influences anglers' substitution behavior. The results showed how much specialization, constraints, and demographic variables are related to anglers' willingness to substitute other fishing locations for one location.

Raymo and Sweeney (2006), investigated relationships between retirement preferences and perceived levels of work-family conflict, evaluated the extent to which work–family conflict was mediating mechanism between stressful work and family circumstances and preferences to retire and explored potential gender differences in the association between work- family conflict and preferring retirement. To achieve this goal, a sample size of 4106 was used with some restrictions about generalization the results to the entire population of the similarly age (52-54 years). The response variable was consist of three categories: working full time, working part time, not working (retired, or something else). Explanatory variables were: family stress spillover into work, work stress spillover into family, potentially stressful job characteristics, potentially stressful family characteristics, marital status and relationship quality, sex, hourly wage, pension eligibility, health insurance, self-rated health, educational attainment, has working spouse. Researchers

evaluated the hypotheses by estimating a series of MLR models, a single model has been estimated for both men and women, initial exploratory analyses indicated that adding the full set of interactions with gender did not significantly improve model fit. They also considered potential violations of the assumption of independence of irrelevant alternatives, but Hausman tests showed no difference in parameters for either full or partial retirement when the other alternative was not available. The response variable in all models was the log odds of preferring either to be working part time or not to be working at all relative to working full time 10 years later (age 62–64). Researchers produced four models, model 1 was the baseline model, and another three models to check three hypotheses were putting by the researchers.

Murray, (2007), used MLR to estimate the effects of individual and community factors on a death. Diabetes were recorded as one of the multiple contributing causes of death (MCD) being assigned to diabetes as the underlying cause of death (UCD) versus assignment to cardiovascular, other non-communicable, or communicable diseases. Data for causes of death were from the National Center for Health Statistics National Vital Statistics System in the U.S. The response variable with four categories and the study dialed with 11 explanatory variables. The study used the multinomial logistic regression to estimate the relative risk ratios (RRRs) of a death, for which diabetes were recorded. The explanatory variables were exogenous individual- and community-level variables were eleven explanatory variables.

Slingerl and etat, (2007), examined the effect of retirement on changes in the three major aspects of physical activity as response variable (work-related transportation, sports, non-sports leisure time). Over 13 years' follow-up among employees aged 40–65 years who participated in the

Globe Study: "Health and Living Conditions of the Population of Eindhoven and surroundings 1991–2004". Specifically, the study hypothesized that people who retired during follow-up would be more likely to have reduced their work-related physical activities but increased their sports and non-sports leisure-time physical activities. The study performed analyses on retired and employed participants for whom complete data on age, sex, marital status, chronic disease, and education were available at baseline (n=971), to make the sample is representative of the original population. Researcher applied weighting factors, taking into account the sampling design and no response. MLR model used to explore the effect of retirement on the change in physical activity between baseline and follow-up. Full models without missing values were used for each of the aspects of physical activity. MLR model indicated that retirement was associated with significantly higher odds for a decline in work-related transportation physical activity. Chronic diseases, and education, remarkably, retirement were not significantly associated with a decline. In sports participation, retirement was associated with significantly lower odds for a decline in non-support's leisure-time physical activity, compared with those remaining employed but not with an increase in non-sports leisure-time physical activity.

Takagi etat. (2007), investigated individual–level conditions and prefecture–level contextual factor that enable and or restrict intergenerational co-residence arrangements between older parents and adult children. To achieve the subject, the researchers used sub group of data from the baseline sample of Nihon University that focuses on the health and social conditions of the population age 65 years and older in Japan. The person had at least one living child (sample size 3565). Response variable had three categories (does not live with children, life-

long co residence, and boomerang co- residence). Explanatory variables covered demographic characters, socioeconomic resources and needs, family composition and availabilities, and normative attitudes (total explanatory variables 13 variables). Sample size (n=3565) distributed as 42% for does not live with children, 45% for–life-long co residence, 13% for boomerang co residence). Instrumental activities were consisting of seven variables each with four points of scales regarding to health status. Researchers found some of these variables were skewness, so they combined all seven items and create a dichotomous variable with the name instrumental activities. Researchers employed the MLR using hierarchical Generalized Linear Model (GLM). They considered non-co residence as baseline category, and they contrasted life–long and boomerang with the baseline category using a logit function in which the likelihood of each was estimated within two level structure. Level 1 for individuals, and level 2 for prefectures. They noted that the statistical approach employed a random effects model which intercept and slope coefficients of individual level predictors varied across prefectures, and were then predicted by prefecture level characteristics. Researchers estimated three models by sequentially including individual level predictors, prefecture level predictors. Finally they explained and interpreted the relationship and effect of each of independent variables in the model.

Brannon et al, 2007, tried to assess how perceived rewards and problems with care giving work and supervision relate to intent to leave among direct care workers who employed in provider organizations participating in the better jobs better care. To achieve this goal, researchers selected a sample from five state completed a paper survey form 139 health organizations, and sample size was 3039 of the employees according to

the criteria. The response variable was the intent of the present job with three scales (not at all likely to leave, very likely to leave in the next year, and somewhat likely to leave next year), the baseline category was the first. The explanatory variables used as: first the sample divided according to kind of the job (four scale levels: skilled nursing (1262), home care (1306), assisted living (425), and adult day services (46). The researchers analyzed this four groups separately, through seventeen explanatory variables: hazards, dead end, overload, discrimination, challenge, recognition, helping others, decision authority, team spirit, income, supervisor quality, job tenure (months), age, non-white race, post high school education, self-efficacy, job alternative. Age was reported in categories but the researchers used it as continuous variable by using the midpoint of each category with higher values indicating increased age. Also the race variable measured by 7 levels and for analysis it was collapsed into dichotomous variable because there were small frequencies in some categories, also the same thing with education levels. Analysis: the researchers ran full MLR on the total sample, and then it ran separately for nursing facility, home care, and assisted living. Regarding to the adult day services it was too small for used to analyze separately so the researchers included in total sample but did not analyze them separately.

Kang, H. (2008) in a study to determine the determinants of academic performance as a qualitative response variable. The study used multinomial logit analysis to identify those variables that determine a student's grade in an undergraduate money and banking course. The results suggest that the key determinants are an adjusted cumulative GPA and per- centile rank on a college entrance exam. Less important are a student's attendance record and the student's overall value of the course.

Generally, effort and intelligence determine the grade. Demographic variables, such as commuting distance, age, sex, and living arrangements, do not seem to con- tribute. Measures of counter effort, such as hours worked on an outside job, do not seem important. Furthermore, measures of preparation, such as prior credit for or concurrent enrollment in suspected complementary courses show no significance relation.

Riggs (2008), explored how the sexuality and parent status of men, and the context in which they donate, were potentially associated with three variables: motivations to donate, understandings of the meanings of biology or genetic material, and the determination of children's best interests. The sample was (30) semi-structured interviews were conducted by the author with Australian gay. Heterosexual men who have acted as known sperm donors. The average age of participants was 45 years, the range being 25 to 65. Distinguishes, of this study from a statistical standpoint was the way followed in the construction of the response variable by using three response variables. First, motivations: within this variable three categories were constructed on the basis of their prevalence across a number of participants. Second: meaning of biology within this variable also identified three categories. Third: determine children' best interests, within this variable, two categories were identified, these three variables were used to consist of the response variable with three categories. Regarding to explanatory variables were the categorical variables, sexuality (gay, or heterosexual), parenting status (whether they currently cared for children on a custodial basis or not). Tests of association were conducted to assess their impact upon the response variable categories. Initial chi square tests performed on the coded data suggested that despite the small sample size, the findings were statistically significant. However, as the use of chi square tests is not indicated for data where more than 10% of the cells have expected

frequencies less than five. Log-likelihood ratio tests were performed. Log-likelihood ratio tests are appropriate for use with small sample sizes that result in cells with expected frequencies less than 5, and where there are more than two levels on the response variable. The findings presented from these tests indicated that the explanatory variables may in combination be associated with each of the response variable, rather than solely as individual isolated variables. The study employed the MLR model to examine the associations between the response variable and explanatory variable. The analysis is amenable not only to small sample sizes, but also to samples where the response variable have more than 2 categories, whilst it has been suggested that MLR model are best conducted on larger data sets. It is nonetheless possible to assess the validity of findings derived using such analyses with small sample sizes. Riggs suggested that small sample sizes should primarily become of concern for MLR model when the standard error presented in the parameter estimates is exceptionally high. An examination of the SE values presented in the parameter estimate tables for each variable in the analysis would indicate that this was not the case in the present study. Also suggested using the Hosmer and Lemeshow chi square test, rather than the standard chi Square test, as this is more appropriate for use with small samples. All chi square tests utilized in this study were thus the Hosmer and Lemeshow method. Finally, Riggs suggested that the validity of MLR model with small sample sizes can be assessed by the log ratio values themselves, where exceptionally high log ratio values would indicate questionable validity of the findings. It is important to note the small sample size and the effect. This may have had on overestimating the significance of the test outcomes and thus the rejection of the null hypothesis. However, it is possible to assess the relative degree of concern that should be granted to this likelihood of the null

hypothesis being incorrectly rejected, by examining the size of the standard error (SE) in each of the parameter estimates.

Moorman and Carr (2008), explored the extent to which older adults accurately report their spouses' end-of-life treatment preferences. In the hypothetical scenarios of terminal illness with severe physical pain and terminal illness with severe cognitive impairment, and investigated the extent to which accurate reports. Inaccurate reports, and uncertain reports were associated with spouses' advance care planning and surrogates' involvement in the planning. Also the study used data from married couples who participated in the Wisconsin Longitudinal Study in 2004. By using 2,750 couples were in their mid-60s and in relatively good health, and MLR was conducted. The goal of the study was to identify the factors that were associated with uncertainty. The study focused on a large sample of healthy, community dwelling, older married couples in order to examine the extent to which discussions, living will completion, and durable power of attorney for health care appointment effect the accuracy of surrogates' assessments of their spouses' preferences.

Further, evaluate the extent to which the effect of planning on surrogate in accuracy persists after the control for surrogates' own treatment preferences, and for demographic, religious, and experiential factors that have been shown elsewhere to be associated with accuracy. The study used two response variables to conduct two separate MLR. Each response variable had four categories. The analysis focused on the 2,750 married couples in which both partners completed the module. First, the researchers conducted one-way analyses of variance with post hoc Tukey tests to evaluate significant differences in the means of the explanatory variables among the four subgroups response variable codes. Then estimated MLR for each scenario to identify the correlates of accurate

assessments (reference category) versus errors of over treatment, errors of under treatment, and uncertain responses. Model 1 showed the effects of spouses' end-of-life planning behaviors. Model 2 was further adjusted for graduate socio-demographic characteristics, religious affiliation, and death attitudes, cognitions, and experiences.

Anass BAYAGA (2010), in his study "Usage and application of multinomial logistic regression in risk analysis" the objective of the study was to explore the usage of multinomial logistic regression (MLR) in risk analysis. In this regard, performing MLR on risk analysis data corrected for the non-linear nature of binary response and did address the violation of equal variance and normality assumptions. Additionally, use of maximum likelihood (-2log) estimation provided a means of working with binary response data. The relationship of independent and dependent variables was also addressed. The data used included a cohort of hundred risk analyst of a historically black South African University. The findings revealed that the probability of the model chi-square (17.142) was 0.005, less than the level of significance of 0.05 (i.e. $p<0.05$). Suggesting that there was a statistically significant relationship between the independent variable-risk planning (Rp) and the dependent variable-control mechanism (control mecs) ($p<0.05$). Also, there was a statistically significant relationship between key risks assigned (KSA) and time spent on risk mitigation. For each unit increase in confidence in control mecs, the odds of being in the group of survey respondents who thought institution spend too little time on Rp decreased by 74.7%. Moreover, the findings revealed that survey respondents who had less confidence in control mecs were less likely to be in the group of survey respondents who thought institution spent about the right amount of time on risk planning.

Mala, A. (2010) carried out a study to assess the relationship between the benzene concentration and trans-trans-muconic acid (t,t-MA), biomarkers in urine samples from petrol filling workers. A total of 117 workers involved in this occupation were selected. The multinomial logistic regression equations were used to predict the relationship between benzene concentration and t,t-MA. The results showed a significant correlation between benzene and t,t-MA among the petrol fillers. Prediction equations were estimated by adopting the physical characteristic viz., age, experience in years and job categories of petrol filling station workers. The study showed that, there was no significant difference observed among experience in years. Petrol fillers and cashiers having a higher occupational risk were in the age group of ≤24 and between 25 and 34 years. Among the petrol fillers, the t,t-MA levels with exceeding ACGIH TWA-TLV level was showing to be more significant. This study demonstrated that multinomial logistic regression is an effective model for profiling the greatest risk of the benzene-exposed group caused by different explanatory variables

Abd alla M. EL-HABIL (2012), in his study aims to identify the application of Multinomial Logistic Regression model in practical way using real data on physical violence against children in Gaza, he found that sex is the most significant explanatory variable on the physical violence against children.

Concerning the rural household income, very little is known about the economics of household level activities in both agriculturalists and pastoral production systems (Etcher and Baker, 1982). Despite this, interventions are frequently proposed which call for increased cash expenditures, the implications of this for different groups of households (poor, rich) needs to be assessed. An intervention may require increased

labour input. Will there be enough labour and if so will sufficient food be available to sustain the energy requirements of increased effort? What is the implication of this for the security and viability of different groups of pastoral and agriculturalists households (rich, poor)? It is in such a context that information on household income and expenditures is required.

Expectations of household income are an outcome of economic conditions and household. Perception of income adequacy has been discussed by psychological economists since Katona first book (1951, 86-112). The concept that the level of satisfaction with income is relative and depends on a reference level of income was introduced by economists over fifty years ago (Duesenberry, 1949). Since then, many authors questioned the mainstream belief that utility depends on absolute income only.

Duesenberry, (1949) stated that the concept that the level of satisfaction with income is relative and depends on a reference level of income was introduced by economists over fifty years ago. Since then, many authors questioned the mainstream belief that utility depends on absolute income only.

Perception of income adequacy has been discussed by psychological economists since Katona first book (1951, 86-112).

Solow (1957) and Nelson (1964) postulated that education adds to the effectiveness of labour through technical progress and however enhance the opportunity of getting more income. In general, education allows people to adapt more easily to both social and technical changes in the economy and, to changes in the demand for labour.

The models of Grossman (1972) and Jacobson (2000) assume that the education of the parents is the most important variable influencing the efficiency of the production process which leads to increase the level of household income.

(Fisher, 1987; Samuelson and Nordhaus, 1992), said that the current studies has adopted the structural approach that links incomes in rural areas with factors that influence production in those areas, including ecological conditions, the size and education of the labour force, and land developments. However, in my side as a researcher I will do add to them the following factors; conflicts, war, poor economic policy and fragile economic background which were represent major factors affected the level of income in the rural areas of Sudan.

Most studies examining the relationship between income and health have used annual family income for the measure of income, as this measure is routinely collected and easy to access. Income level is associated with almost every indicator of health, including infant and adult mortality, morbidity, disability, health behaviors, and access to health care. Individuals in poverty have the worst health, though even people in middle income levels have worse health than people in the highest income level. Low income is associated with many other factors contributing to poor health outcomes, including risky health behaviors, lower levels of education, substandard housing, food insecurity, and lack of health insurance coverage.

On the other hand the “structural or economic” strand argues that most poverty can be traced back to structural factors in an economy or institutional environments that favour certain groups over others. For example, economic opportunities may vary markedly between different

locations with significant impact on income levels and poverty. The poverty of an individual cannot, therefore, be solely attributed to personal characteristics without paying attention to the circumstances prevailing where a person lives (Holzer, 1991). This direct relationship between poverty and income supports the argument that productive work is the best mechanism for lifting people out of poverty which, in turn, suggests that strategies to expand economic opportunities and promote income growth are necessary for sustained poverty reduction.

(Rodgers and Rodgers 1993, pp.44-45), stated that since parents from chronically poor households are, on average, less educated than parents from not-poor households they can be assumed to be less efficient in producing child health.

In their analysis of poverty in Uganda, Okurut et al. (2002) found that the higher the educational attainment of the household head the wealthier the household, while the larger the household size the poorer the household.

In social study, Mung'ong'o and Mwamfupe (2003) found among migrant Maasai pastoralists of Morogoro and Kilosa districts in Tanzania, communities adapt to new conditions; in particular, environmental and political factors have caused a switch in traditional livelihoods from nomadic pastoral activities to sedentary agricultural activities. This has not, however, made a noticeable impact on income and poverty alleviation as the types of activities are simply for subsistence.

Cutler and Lleras-Muney (2006) conclude that higher levels of education do in fact lead to different choices of income sources (i.e. more educated persons were more likely to have a job opportunity).

Sen and Palmer-Jones (2006) examined the link between poverty and location in rural India and concluded that being poor or rich was strongly

related to where a person lived. The study applied spatial econometric methods, including ordinary least squares (OLS) and maximum likelihood (ML) estimation techniques, to explore the determinants of rural income poverty in relation to agricultural growth. The study found that low incomes and poverty were highly correlated with agricultural performance. Therefore, factors that promote agricultural production are important to poverty alleviation and to increase the level of income.

In Argentina, a study on the rural poor found that the principal causes of poverty were low education, poor health facilities and inadequate infrastructure (Verner, 2006).

Kessy and Urio (2006) have shown that the provision of loans by micro-finance institutions boosted the livelihoods of poor Tanzanian households. The study also found that the lack of infrastructure, especially rural roads, was the main reason why micro-finance institutions failed to operate in rural areas and did not make a difference in their level of income.

Places where conditions were unfavorable to the development of irrigation facilities experienced little agricultural growth and, consequently, these areas were characterized by low income levels and minimal declines in rural poverty. These results are supported by Son (2007) who looked at the effect of irrigation and water availability on rural incomes in Vietnam. The study found that between 2000 and 2005 rural incomes more than doubled in irrigated areas compared with non-irrigated areas.

Smith (2007) in a study on the determinants of Soviet household income found that human capital and demographic factors were the main determinants of income. The well-educated, middle-aged and self-

employed people had relatively comfortable incomes, the study also concluded that location had strong influence on household incomes.

On a study regarding the relationship between level of income and child health, researcher had found that, poor households are able to purchase fewer and lower quality market goods and services, including those involved in the production of child health, such as nutritious foods, sporting equipment and medical care. By comparison, families with higher incomes are less constrained in their potential health investment as a smaller portion of their budget must be allocated to other necessities, (Chia (2008, p.233).

M. T. Parvin1 & M. Akteruzzaman (2012), in their statistical study (Factors Affecting Farm and non-Farm Income of Haor Inhabitants of Bangladesh) aims to examine the factors influencing farm and non-farm income of Haor economy in Bangladesh. Dingaputa Haor area of Netrokona district was selected for the study and a sample of 60 farmers had been taken randomly. The log linear form of Cobb-Douglas production function was chosen to determine the effects of socio-economic variables on farm income and non-farm income. Apart from this, some descriptive statistical analysis were done to examine the socioeconomic characteristics of sampled households. The estimated results of the regression models revealed that family size and farm size had a significant positive effect on farm income and non-farm income had a significant negative effect on farm income. On the other hand, family size had a positive and significant effect on non-farm income and farm income had a negative and significant effect on non-farm income. To promote the farm and non-farm sector income and strengthening its potential linkages between them, the study mainly recommends increasing efforts on two fronts: first, reforming the institutions

responsible for rural development and second, development activities and projects that would enhance farm and non-farm income and the linkages between them. The study stated that socioeconomic background and characteristics of the respondents have a vital role in farm and non-farm activities to a great extent. In addition, these characteristics can be used as important indicators in making comparison among different categories of the respondents. A number of socioeconomic aspects of the sample households were examined. These were age, family size, farm size, occupational structure, educational attainment for the members of selected households, farm and non-farm income, employment opportunities etc. Besides that, family size was measured by taking into consideration all the existing family members of the respondent households. In this study, family size was assumed to affect the households' farm and non-farm income. The regression coefficients of family size show that increase in family size would lead to increase in the farming status of the household.

Dayal Talukder, (2014) also conducted a regression study to assess the determinants of Income of Rural Households in Bangladesh, the purpose of the study was to investigate the determinants of income and growth in income of rural households in Bangladesh in the post-liberalisation era. Using data mainly from secondary sources, the study applied the ordinary least square (OLS) regression models to assess the determinants. The determinants were justified based on both initial (1985-86) and current (2005) endowments (household characteristics) for a comparative analysis. The study used both economic and non-economic characteristics simultaneously for considering their joint effects on household income. The regression models revealed that household size was the only non-economic factor that was statistically significant and positive determinant

of household income in both 1985-86 and 2005. Household size was the largest positive determinant and small farmer dummy variables was the largest negative determinant of income in 1985-86. Similarly, household size was the largest positive determinant and farm-household dummy variable was the largest negative determinant of income in 2005.

From the literature reviewed and the previous studies being shown in chapter one, it is clear that logistic regression is widely used in the field of medicine and social sciences which revealed a huge gap in the uses of the method of the multinomial logistic regression in the field of economics especially in term of household wellbeing. However, I would expect that this study will add more and cover that gap in term of the determination of the main factors that would affect the sufficiency of the household income.

1.9 Structure of the study

This research study consists of five chapters. The first chapter is the introduction, which include: the problem of the study, the importance of the study, the objectives of the study, the hypothesis of the study, the methodology of the study, and the previous studies. The second chapter gives conceptual background about South Darfur and the determinants of household income sufficiency. The third chapter is devoted to the methodology of the study and the multiple logistic regression. In chapter four the analysis of the data and discussion of the results are presented. Chapter five, is concerned with the main findings and the recommendations of the study.

Chapter Two

Economy of south Darfur and determinants of income

2.1 Preface

2.2 Economy of south Darfur

2.3 South Darfur economic declining

2.1 Preface

South Darfur State lies between latitudes 11.6489 N and longitudes 24.9042 E, Covered an area of about 49,151 mi², borders country of south Sudan to the south and central Africa to the west in addition to two states north Darfur to the north and eastern Darfur to the east. It has a population of about 2,890,000 (2006) distributed over twenty one localities, namely Kass, Edd Alfursan, Buram, Kabom, Mershing, Gerieda, Nyala Shamal, Nyala Wasat, Alsalam, Alwohda, Netega, Katela, Tulus, Rehid Alberdi, Um Dafoug, Alradom, Alsunta, Beliel, Demso, Sharg Algabal and Shataia. Nyala is the most populated locality (Central Bureau of Statistics, Sudan 2003; 2008). More than 70% of the population is rural and agriculture is the main source of livelihood for residents.

Almost more than 70% of the households in the south Darfur state as a part of the whole Darfur region, depend on agriculture and livestock for their livelihoods. Traditional rain-fed agriculture is the dominant seasonal farming activity across the state. Millet is the main staple food cultivated in the northern and eastern parts of the region while sorghum is cultivated in the south and in the lowlands (wadi). Livestock rearing among the agro-pastoralist groups has considerably diminished due to the conflict that erupted in 2003.

Most of the households tend to keep only a few domestic goats to avoid looting, which is common amongst large herd owners. For agro-pastoralists, the hunger season occurs during the rains between late June and late September when labour requirements are highest but food availability is the lowest. Nearly all households attempt to diversify their incomes by engaging in petty trade, firewood and grass collection and sale, domestic labour, long-distance labour migration as well as to

augment through remittances, gathering and consumption of wild foods. As a result of the current conflict, the disruption of households' livelihoods and coping mechanisms and subsequent displacement for many has contributed to increased food insecurity, lack of the main sources of income, disruption the health situation as well as led to high loses of education opportunity for the young people. WFP (2011), "comprehensive food security assessment"

2.2 South Darfur economy

The economics of Darfur based mainly on agriculture and grazing beside the trade locally and across borders. The diversity of the natural environment resulted in diversification of economic activities and people livelihood. Access to natural resources considered as the main source of wealth, and also represent crucial issue in the daily life and help creating opportunities for the majority of the people.

The activity of the rural economy in the south Darfur state has remained as an economy of living, but a remarkable transition occurred in the eighties of the twentieth century, where production began to exceed the self-sufficiency and become moving towards export and domestic and foreign markets , as a result of that many producers moved to cultivate cash crops like sesame, peanuts, Arabic gum and Kerkade, as well as looking after livestock not only to meet the needs of the living, or to feed the local market but also for export to overseas markets.

Since the conflict began in south Darfur the situation become harmfully and was accompanied by mass displacements for the majority of the rural population of farmers to the main cities, which was reflected directly on the economic situation in the state. The living economy of the majority of the population had become the first victim which led to significant shifts in the base structure of the economic as consequences of the war, and

directed about a third of the population of the state from being producers to be fully displaced. However, this situation not only collapsed the production system, but also creating high rate of unemployment in a shortest time.

2.2.1 Livestock

Greater Darfur region known for a long time as it produces the most important sources of livestock in Sudan, either for domestic consumption or for exportation. A country report for Sudan done by the World Bank (1992) has revealed that the contribution of the livestock trade in the Sudan 's foreign trade surge of 13 % in the seventies of the twentieth century to 50% in the eighties of the same century.

According to the Federation of Chambers of Commerce in Khartoum, the Darfur region was contributing 30 % of the livestock trade before the conflict and fell to less than 15%.

Mohammed Suleiman (2006, p. 362) noted that the outbreak of the conflict in the Darfur region has led to severe effects on the livestock sector, began by large looting for the livestock in the early years of the conflict which led to direct negative impact on livestock producers and on their trade which has become a very difficult task in light of the deterioration of security in the state

Livestock trade has seriously affected by armed and its market is highly collapsed due to the conflict and the owners of the livestock became exposure to the looting network, this situation led them to release their livestock and selling it in a cheap prices which led to the impoverishment of small producers. In addition to that cattle traders also became too vulnerable to looting and then went out of the markets.

According to Gerald (2011) that between 40 -50 of livestock traders agents in Nyala market before the conflict has reduced to become only 10

traders agents of livestock by the year 2011 and has led to a significant impact on the livestock trade and shrinking the size of livestock trade during the years of the conflict by not less than 50% of its size before the conflict.

2.2.2 South Darfur industrial sector

Concerning the industry sector, statistic survey conducted by the Ministry of Industry before the start of the conflict indicates that Darfur was the second- largest industrial area in Sudan after Khartoum. Industrial activities in Darfur are formed mainly from the traditional small industries related to agricultural industrialization, and because most of the traditional industries in Darfur focus on the relevant sectors of the agricultural domain, it has been adversely affected by the terrible deterioration experienced by the agricultural sector in the years of the war and led to an almost complete halt to the industrial sector with the presence of other factors that have also affected the sector led to addition decline, including the decline in electrical supply, low banking finance , the high cost of transportation in addition to marketing difficulties due to weak purchasing power of the citizen.

2.3 South Darfur economic declining

The conflict made significant shifts in the activities of the markets in Darfur. Gerald (2011) monitored that 12 of the biggest markets covered the whole region have been diminished.

The traditional agriculture before the outbreak of armed conflict constitutes a major livelihood for the majority of the population of rural Darfur and represent main source of subsistence economy, as well as the contribution of the cash crops production as a source of income. However, due to the nature of agriculture which need stable situation the

conflict forcing most farmers to leave their lands looking for security and safety and then hindered their abilities for production.

Displacement led to a significant deterioration in the agricultural sector due to reduced plantings area in Darfur from 10 million acres to less than 4 million acres a rate of 60% with reference to the area cultivated in Darfur represents about 25% of the total arable land in Sudan, amounting to an average of about 40 million acres annually.

With respect to the contribution of the agricultural sector in the GDP to Darfur for the period prior to the conflict (2000 - 2004), a study for the World Food Programme (2006) indicates that the crops were constitute an essential source of wealth and it is contribution in the GDP range between 63 % to 75%. And it has fallen by 50 % during the first two years of the start of the conflict and livestock fell by 9 % during the first year of the conflict. As a result of that deterioration of the agricultural sector caused serious decline in the food security, and then the people in the region were shifted from a self -sufficient food crop production to become looking for humanitarian assistance and food aid.

Smith " 2011 " refers to a report on " the impact of the conflict on trade and markets in Darfur," said that many traders have been forced to abandon their trade at the beginning of the conflict because of insecurity and displacement, and that those who were able to continue their businesses found themselves languishing under the weight of additional taxes imposed by the state government to compensate for losses incurred due to the decline in the business activities. In addition to the spread of the "shadow economy" which is not subject to government control and the emergence of the phenomenon of parasitic economic activities and appearance of an unproductive businesses as an alternative for those who lost their source of livelihood.

Furthermore, the subsidies and aids provided by a number of non-governmental and voluntary organizations that have proliferated in the region in order to bridge the food gap caused by leaving the citizens for their sites of agricultural production due to displacement have contributed significantly in distorting the consumption pattern of a large number of the community in Darfur, and it has become one of the most reasons that prompted citizens to rely completely on subsidies and move away from the self-production to be dependent, and on other side most of them have make it as a source of income generation through selling it in a ranged markets.

South Darfur also has suffered from the unemployment problem because of the war that have hindered real income growth per capita. According to Kloiber (2007), the economy of conflict lose about 2.5 percentage points of growth per year in comparing with the economy in peace situations. Beside that the problem of unemployment is increased due to the increasing population growth and the traditional economic sector failed to absorb the new labor force.

Also poor infrastructure in the state especially the paved road that links the state with the center and marketing areas in addition to the energy savings and limited water are all factors that played a big role in making the situation more deteriorated and weakened the economic potential of the state.

2.4. Income sources details (income determinants)

Sale of cereals (sorghum, millet), sale of other crops, sale of livestock and animal products, remittances, renting out donkey cart, gifts from family/relatives, sale of food aid, agricultural wage labor, salaried work, skilled labor, wheel barrow/trolley, domestic labor, brick-making, construction, pottering, sale of water, tea seller/catering, rickshaw driver,

sales of handicraft, sales of firewood/grass, sale of charcoal, other petty trade, others. WFP (2011), “comprehensive food security assessment”

As in the above mentioned paragraphs, agriculture and livestock in addition to trade are representing most important sources of income to the whole societies in the state. The ongoing conflict in the region for more than a decade considered as one of the most important reasons that led a majority of the people to lose their main sources of income then weakening their ability to meet the basic requirements and influence the adequacy their income.

Chapter Three

Multiple Logistic Regression

3-1 Preface

3-2 Logistic Regression Model

3-3 Multiple Logistic Regression

3-4 Multinomial Logistic Regression Model (MLR)

3-5 Assumptions of the MLR model

3-6 Strategies in model selection

3-7 Diagnostic of MLR model

3-8 Model validation

3.1 Preface

Logistic regression is used increasingly in a wide variety of applications. Early uses were in biomedical studies but the past 20 years have also seen much use in social science research and marketing. Recently, logistic regression has become a popular tool in business applications. Some credit-scoring applications use logistic regression to model the probability that a subject is credit worthy. For instance, the probability that a subject pays a bill on time may use predictors such as the size of the bill, annual income, occupation, mortgage and debt obligations, percentage of bills paid on time in the past, and other aspects of an applicant's credit history, Agresti (Analyses of Categorical data, 2002).

Logistic regression is a promising statistical technique that can be used to predict the likelihood of a categorical outcome variable. It has found widespread use in the epidemiological literature, where often the dependent variable is presence or absence of a disease state. This technique has also proven useful in broader areas of social sciences (e.g., Chuang, 1997; Janik and Kravitz, 1994; Tolman and Weisz, 1995) and education, especially higher education (Austin, Yaffee, and Hinkle, 1992, Cabrera, 1994; Peng, So, Stage, & St. John, 2002) than the typical epidemiological situation.

Logistic Regression technique yields coefficients for each independent variable based on a sample of data (Huang, Chai and Peng, 2007). Logistic regression models (LRM) with two or more explanatory variables are widely used in practice (Haines and Others, 2007). The parameters of the logistic regression model are commonly estimated by maximum Likelihood, Pardo (2005). The advantage of logistic regression is that, through the addition of an appropriate link function to the usual

linear regression model, the variables may be either continuous or discrete, or any combination of both types, and they do not necessarily have normal distributions (Lee, 2004). The predictor values from the analysis can be interpreted as probabilities (0 or 1 outcome) or membership in the target groups (categorical dependent variable). It has been observed that the probability of a 0 or 1 outcome is a non-liner function of the logit (Nepal, 2003). Logistic Regression is similar to a linear regression model but is proficient to models where the dependent variable is dichotomous. Logistic Regression coefficients can be used to estimate odd ratios for each of the independent variables in the model. Logistic Regression helps to form a multivariate regression between a dependent variable and several independent variables (Lee, Ryu and Kim, 2007). It is designed to estimate the parameters of a multivariate explanatory model in situations where the dependent variable is dichotomous, and the independent variables are continuous or categorical.

The review of the evolution of the logistic regression model over the last century cannot be dissociated from the development of methods of analysis of categorical data and its tools. In fact, this development has been associated with the development of these methods particularly the Maximum likelihood (ML), chi-squared distribution, and odds ratio (OR) concept. Alan Agresti in his famous book" Categorical Data Analysis" believes that the year of 1900 is an apt starting point of the history of categorical data analysis (CDA).

There are multiple ways to describe the mathematical model underlying multinomial logistic regression, all of which are equivalent. This can make it difficult to compare different treatments of the subject in different texts. The article on logistic regression presents a number of equivalent

formulations of simple logistic regression, and many of these have equivalents in the multinomial logit model.

The idea behind all of them, as in many other statistical classification techniques, is to construct a linear predictor function that constructs a score from a set of weights that are linearly combined with the explanatory variables (features) of a given observation In recent years, specialized statistical methods for analyzed categorical data have increased, particularly for application in biomedical and social science. Regression analysis is one of these statistical tools that utilize the relationship between two or more variables. It is the name of a set of techniques that attempts to predict one variable called response variable from another variable, or se t of variables called explanatory variables or predictor variables. This tool can serve three major purposes: prediction, explanation, control, as the regression equation can be used to predict an individual's score on the outcome variable of interest, and can explain why certain events occurred, based on their relationship, and to control for other variables. The regression models can be divided into two groups, the first related to linear relationship models, and the second related to non-linear relationship models. The linear models, considered up to this point, are satisfactory for most regression applications. Nonlinear model used when the linear model is not suitable anyhow. The model is considered a linear if the parameters can presented by linear relationship, and it is not necessary include first order model but also it can be more complex models. Models can handle more complicated situations, such as analyzing simultaneously the effects of several explanatory variables. In the same time a good-fit model has several benefits like structure form of the model that describes the patterns of association, interaction between variables, determines the strength and

importance of the effects of the size of the model parameters. A goodness-of-fit can be inferences about the parameters, evaluates which explanatory variables effect the response variable, while controlling effects of possible confounding variables, and the model's predicted values smooth the data and provide improved estimates of mean of response variable at possible explanatory variable values. Many of statisticians believe that the logistic regression model is one of the important models can be applied to analyze a categorical data, this model is our interest in this thesis, and to understand the basic ideas of this model. It is necessary to introduce Generalizes Linear Model (GLM) theory, as logistic regression model is a special case of GLM.

3.2 Generalized Linear Model on Categorical Data Analysis (CDA)

The theory and an algorithm appropriate for obtaining maximum likelihood estimates where the response follows a distribution in the exponential family was introduced in 1972 by Nelder and Wedderburn, They introduced the term GLM to refer to a class of models that could be analyzed by a single algorithm. The theoretical and practical application of GLMs has received attention in many articles and books. GLMs contain a wide range of models such as linear regression, binary logistic regression, and Poisson regression for count data outcomes. The GLM requires a link function that characterizes the relationship of the mean response to a vector of covariates. In addition, a GLM requires of a variance function that relates the variance of the outcomes as a function of the mean. The derivation of the iteratively reweighted least squares algorithm appropriates for fitting GLMs begins with the likelihood for the exponential family.

3.2.1 The Components of Generalized linear model (GLM)

Let Y is the response variable with probability distribution, and it has the observations $(Y_1, Y_2, \ldots Y_n)$. Standard GLM treats $(Y_1, Y_2, \ldots Y_n)$ as independent. GLM have three components, the Random component which is identified the response variable Y, the Systematic component which is specified the explanatory variables of the model, and the Link function which is specified a function of expected value of Y. In many applications the observations on Y are binary (success or failure), more generally, each Yi might be the number of successes out of a certain fixed number of trials. In either case, we assume a binomial distribution for Y. In some applications, each observation is count; we might then assume a distribution for Y that applies to all the nonnegative integers, such as the Poisson or negative binomial. If each observation is continuous, such as subject's weight, we might assume a normal distribution for Y. The systematic component specifies the explanatory variables for the model, these enter linearly as predictors on the right hand of the model equation, variables that are the set of (X_j) in the formula: $\alpha + \beta_1 X_1 + \ldots + \beta_k X_k$ This linear combination of the explanatory variables is called "linear predictor". GLM use lower cases for each x to emphasize that x –values are treated as fixed rather than as random variable. Link function specifies a function of expected value (mean) of Y, which the GLM relates to explanatory variables though a prediction equation having linear form. Denote the expected value of Y, the mean of its probability distribution by $\mu = E(Y)$, this component specifies a function g (.) that relates μ to the linear predictor as:

$$g(\mu) = \alpha + \beta 1 X 1 + \ldots + \beta k X k \qquad (3.1)$$

g (.) link function, it connects the random and systematic components. The simplest link function is, g (μ) = μ, called identity link, it specifies a linear model for the mean response $\mu = \alpha + \beta 1X1 + ... + \beta kXk$ this form of ordinary regression model for continuous responses. Other link function permits μ to be nonlinearly related to predictors like a link

function g(μ) = log (μ), this link function models is the log of the mean, the log function applies to positive numbers, so the log link function is appropriate when μ cannot be negative, such as with count data. A GLM that uses the log link called loglinear model it has the form

$$\log(\mu) = \alpha + \beta 1X1 + ... + \beta kXk \qquad (3.2)$$

Another link function like $g(\mu) = \left[\frac{\mu}{1-\mu}\right]$, this link function model, log of an odds, it is appropriate when μ is between 0 and 1, such as a probability, this is called the logit link. A GLM that use the logit link is called a logistic regression model. Each potential probability distribution for Y has one special function of the mean that is called natural parameter. For normal distribution, it is the mean itself. For the binomial the natural parameter is the logit of the success probability. Early analysis of non-normal responses often attempted to transform Y so it is approximately normal, with constant variance, then ordinary regression methods using least squared are applicable. With the theory and methodology of GLM it is unnecessary to transform data so that methods for normal response apply, because the GLM fitting process uses maximum likelihood methods (ML) for the choice of random component whereas not restricted to normality for that choice, also the choice of link function is separate from the choice of random component, it is not chosen to produce normality or stabilize the variance, Agresti (2007).

3.2.2 Generalized linear model (GLM) for Binary data

Many categorical response variable Y have only two categories, we denote the two possible outcomes, Success "1" and Failure "0". The distribution of Y is specified by probabilities for one outcome is P(Y=1) = π of success, and P(Y=0) = (1- π) of failure, and its mean is E (Y) = π For n independent observations, the number of success has binomial distribution, specified by n and π , the formula is bin (n, π), each binary observation is a binomial variable with n = 1 has expected value E(y) = n π, and variance of (y) = n π (1- π).

GLMs can have multiple explanatory variables, for simplicity, we introduce them by a single x, in this case, the value of π can vary as the value of x changes, and we replace π by π (x) when we want to describe its dependence on that value.

3.3 Logistic Regression Model

Relationships between π (x) and x are usually non-linear rather than linear, a fixed change in x may have less impact when π is near 0 or 1 than when π near the middle of its range. In practice, π (x) often either increases continuously or decreases continuously as x increases, the S–shaped curves are often realistic shapes for the relationship.

The most important mathematical function with this shape has formula:

$$g(x) = \frac{\exp(\alpha + \beta x).}{1 + \exp(\alpha + \beta)} \qquad (3.3)$$

Using the exponential function, this is called the logistic regression function, so the logistic regression model is

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x \qquad (3.4)$$

This model is called logistic regression model, and it is a special case of GLM, random component for (success, failure), outcome has a binomial distribution, the likelihood function is the logit function. The logit function $\log[\pi/(1-\pi)]$, and logit(π) (logistic regression model) are often called logit model. Whereas, π is restricted to 0-1 range, and the logit can be any real number with the potential range for linear predictors (such as $\alpha + \beta$ x), the parameter $\beta > 0$ then π (x) increases as x increases β and if $\beta < 0$, then π (x) decreases as x increases, so β determines the rate of increase or decrease of the curve, if $\beta = 0$, the curve flattens to a horizontal straight line.

3.3.1 Probit Regression Model

Another model that has S–shaped curve is called the probit model, the link function is called probit link, transforms probabilities to z scores from the standard normal distribution. Probit model has expression: $probit[\pi(x)] = \alpha + \beta x$, the probit link function that applied to π (x) gives the standard normal z-score at which the left-tail probability equals π (x). In practice, probit and logistic regression models provide similar fits. If a logistic regression model fit well then so does the probit model, and conversely. Anyhow, logistic regression was not developed until the mid 1940, and not used much until 1970. It is now more popular than the probit model. Also logistic regression model parameters related to odds ratios, thus one can fit the model to data from case control studies, because one can estimate odds ratios for such data. Finally, the development of GLM theory in the mid-1970s unified important models

for continuous and categorical response variables, a nice feature of GLM is that the model-fitting algorithm, Fisher scoring, is the same for any GLM, this holds regardless of choice of distribution for Y or link function, Agresti (2007).

3.3.2 Loglinear Analysis

Loglinear analysis is used when all the interested variables are categorical, and the objective is to find out which one of the interactive relationships among the variables can best explain the observed frequencies rather than explain one variable's variation with other variables. Because data with categorical variables can be presented in a contingency table, loglinear analysis is sometimes called multi-way frequency analysis. It is also a multivariate version of chi-square analysis, dealing with variations and interactions between three or more categorical variables. In loglinear analysis, none of the variables is treated as a response variable; rather it is the cell count that is to be explained. Logistic regression should be used if a response variable is assigned and it is categorical. Loglinear analysis became a constraint when both categorical and continuous variables included in a model. Logistic models can handle both type of variables as explanatory variables (only the response variable must be category variable), Salkind (2007).

3.3.3 Multiple Logistic Regression

The logistic regression can be extending to models with multiple explanatory variables. Let k denotes number of predictors for a binary response Y by $X_1, X_2,.,.,., X_k$ the model for log odds is;

$$Logit\left[P(Y=1)\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \qquad (3.5)$$

and the alternative formula, directly specifying $\pi(x)$, is

$$\pi(x) = \frac{\exp(\alpha + \beta_1 X_1 + ... + \beta_k X_k)}{1 + \exp(\alpha + \beta_1 X_1 + ... + \beta_k X_k)} \quad (3.6)$$

The parameter βi refers to the effect of Xi on the log odds that Y =1, controlling other Xj, at fixed levels of other Xj.

3.4 The multiple logistic regression (MLR) model

The MLR model is generally used where the response variable is composed of more than two levels or categories. The basic concept was generalized from binary logistic regression. Continuous variables are not used as response variable in logistic regression, and only one response variable can be used. MLR can be used to predict a response variable on the basis of continuous and/or categorical explanatory variables to determine the percent of variance in the response variable explained by the explanatory variables; to rank the relative importance of independents to assess interaction effects; and to understand the impact of covariate control variables. MLR allows the simultaneous comparison of more than one contrast, that is, the log odds of three or more contrasts are estimated simultaneously, Garson (2009). The logistic regression model assumes that the categorical response variable has only two values, in general, 1 for success and 0 for failure. The logistic regression model can be extended to situations where the response variable has more than two values, and there is no natural ordering of the categories.

Natural ordering can be treated as nominal scale, such data can be analyzed by slightly modified methods used in dichotomous outcomes, and this method is called the multinomial logistic. If we have n

independent observations with p - explanatory variables, and the qualitative response variable has k categories, to construct the logits in the multinomial case, one of the categories must be considered the base level and all the logits are constructed relative to it. Any category can be taken as the base level, so we will take category k as the base level. Since there is no ordering, it is apparent that any category may be labeled k. Let πj denote the multinomial probability of an observation falling in the jth category, to find the relationship between this probability and the p explanatory variables, $X_1, X_2, ..., X_p$, the multiple logistic regression model then is;

$$\log\left[\frac{\pi_j(x_i)}{\pi_k(x_i)}\right] = \alpha_{oi} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \ldots + \beta_{pj}x_{pi} \qquad (3.7)$$

Where j = 1 , 2 , . . , (k - l), i = 1 , 2 , .., n. Since all the π's add to unity, this reduces to;

$$\log(\pi j(x_i)) = \exp(\underline{\alpha_{oi} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \ldots + \beta_{pj}x_{pi}})$$

$$1 + \sum_{j=1}^{k-1} \exp(\alpha_{oi} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \ldots + \beta_{pj}x_{pi}) \qquad (3.8)$$

for j = 1,2, . . ,(k-1), the model parameters are estimated by the method of ML. Practically, we use statistical software to do this fitting, Chatterjee and Hadi (2006).

3.4.1 Baseline-Category Logit Model

In MLR model, the estimate for the parameter can be identified compared to a baseline category. We defined bold letter as matrix or vector, let $\pi_j(x) = p(Y = j|x)$ at a fixed setting x for explanatory variables, with

$\sum \pi_j(x) = 1$, for observations at that setting, we treat the counts at the J categories of Y as multinomial with probabilities, $\{\pi_1(x),\ldots,\pi_J(x)\}$, logit models pair each response category with a baseline category, often the most common model is:

$$\log\left[\frac{\pi_j(x)}{\pi_J(x)}\right] = \alpha_j + \beta'_j x \qquad (3.9)$$

where j = 1,…, (J −1), simultaneously describes the effects of x on these (J-1) logits, the effects vary according to the response paired with the baseline, these(J-1) equations determine parameters for logits with other pairs of response categories. Since;

$$\log\left[\frac{\pi_a(x)}{\pi_b(x)}\right] = \log\left[\frac{\pi_a(x)}{\pi_J(x)}\right] - \log\left[\frac{\pi_b(x)}{\pi_J(x)}\right] \qquad (3.10)$$

categorical predictors, Pearson chi-square statistic χ^2 and the likelihood ratio chi-square statistic G^2 goodness-of-fit statistics provide a model check when data are not sparse. When an explanatory variable is continuous or the data are sparse, such statistics are still valid for comparing nested models differing by relatively few terms, Agresti (2002).

3.4.2 Estimating multinomial response probabilities

The equation that expresses multinomial logit models directly in terms of response probabilities $\{\pi_j(x)\}$ is;

$$\pi_j(x) = \frac{\exp(\alpha j + \beta' j x)}{1 + \sum_{h=1}^{J-1} \exp(\alpha j + \beta' j x)} \qquad (3.11)$$

with αj = 0 and β J = 0. This follows from;

$$\log\left[\frac{\pi_j(x)}{\pi_J(x)}\right] = \alpha_j + \beta'_j x \quad , \quad j=1,\ldots,(J-1) \qquad (3.12)$$

using the fact that (3.15) also holds with j= J by setting αj = 0 and β J = 0, also, the parameters equal zero for a baseline category for identifiability reasons, the numerators for various j sum to the denominator, so $\sum \pi_j(x) = 1$, for (J = 2), (3.15) simplifies to the formula used for binary logistic regression, Agresti (2002).

3.4.3 Assumptions of the multiple logistic regression model

Using of logistic regression enables to overcome many of the restrictive assumptions of OLS regression. Logistic regression does not require linear relationships between the explanatory variables and the response variable. It does assume a linear relationship between the explanatory variables and the log odds (logit) of the response variable. One strategy for mitigating lack of linearity in the logit of a continuous covariate is to divide it into categories and use it as a factor, thereby getting separate parameter estimates for various levels of the variable. The response variable does not need to be normally distributed but it does assume that distribution is within the range of the exponential family of distributions. Homogeneity of variance assumption does not need. Normally distributed error terms are not assumed. The model should try to inclusion

of all relevant, because, if relevant variables are omitted, the common variance they share with included variables may be wrongly attributed to those variables, or the error term may be inflated. The model should exclusion of all irrelevant variables to avoid; the common variance they share with included variables; may be wrongly attributed to the irrelevant variables; also the more the correlation of the irrelevant variable(s) with ot her explanatory variables; the greater the SE of the regression parameters for these explanatory variables. The error terms are assumed to be independent, violations of this assumption can have serious effects represented in the multicollinearity problem. The problem of multicollinearity will occur in logistic regression, as it does in OLS regression, as the explanatory variables increase in correlation with each other, the standard errors of the logit (effect) parameters will become inflated. Multicollinearity does not change the estimates of the coefficients, only their reliability.

3.5 Building of Multiple logistic regression model

3.5.1 Model selection

A simple model that fits adequately has the advantages of model parsimony. Model parsimony states that the fewer variables used to explain a situation, the more probability that the explanation will be closer to reality, which confers more generalizability on the explanation, Salkind (2007). If a model has relatively little bias, describing reality well, it tends to provide more accurate estimates of the quantities of interest as another criterion can be used additional to significance tests. To select a good model in terms of estimating quantities of interest, it can be used Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). Both judges a model by how close it fitted values tend

to be the true values. In terms of a certain expected value, even though, a simple model is farther from the true model than is a more complex model, it may be preferred because its tends to provide better estimates of certain characteristics of the true model. We will reviews the both criterions with some details: AIC is a way of model selection criterion. It was introduced in 1973 by Hirotogu Akaike as an extension to ML principle. Conventionally, ML is applies to estimate the parameters of a model, Salkind (2007). AIC can be defined as the following: consider (M_1, M_2, … ,M_L) denoted of candidate family of models, let $\theta_k (k = 1,2,..., L)$ denote to the parameter vector for model M_k, and let d_k denote the dimension of the model M_k that is, the number of functionally independent parameters in θ_k , let $L(\theta_k|y)$ denote the likelihood for θ_k based on the data y, and let $\hat{\theta}_k$ denote the ML estimate of θ_k. The AIC for model M_k is defined as:

$$AIC_k = -2 \log L(\hat{\theta}_k|y) + 2d_k \qquad (3.13)$$

The minimum AIC procedure is employed as follows: selecting the fitted model corresponding to minimum value of AIC, one is hoping to identify the fitted model that closest to the generating model. Another illustration, AIC judges a model by how close its fitted values tend to be to the true expected values, as summarized by a certain expected distance between the two. The optimal model is the one that tends to have its fitted values closest to the true outcome probabilities, this is the model that minimizes

AIC = − 2 (log likelihood −number of parameters in model). (3.14)

BIC is a statistic used for comparison and selection of statistical models. BIC is given by a simple formula that uses only elements of standard

output for fitted models. It is calculated for each model under consideration. Models with small values of BIC are preferred. BIC formula and with the smallest value are motivated by one approach to model selection in Bayesian statistical inference. Suppose D a set of data of size n, with statistically independent observations and the "effective sample size" in some appropriate sense when the observations are not independent. Suppose that alternative models M_k are considered for D, and that model is fully specified by a parameter vector θ_κ with P_k parameters. Let $p(D|\theta_k, M_k)$ denote the likelihood function for model M_k, $I(\theta_k) = \log\ p(D|\theta_k, M_k)$ the corresponding log-likelihood, and $\hat{\theta}_k$ the ML estimate of θ_k , let *Ms* denote a saturated model that fits the data exactly, one form of the BIC statistic for a model M_k is

$$BIC_k = -2\left[I(\hat{\theta}_k) - I(\hat{\theta}_s)\right] - df_k \log n = G_k^2 - df_k \log n \qquad (3.15)$$

Where $I(\hat{\theta}_s)$ is the log-likelihood for the saturated model G_k^2 is the deviance statistic for model M_k and df_k is its degrees of freedom, Salkind (2007).

3.5.1.1 Selection of response variable

MLR support only a single response variable with more than two categories. It is important to be careful to specify the desired reference category, which should be meaningful, MLR by default predicts all categories of the response variable except one category which is used as a reference category.

3.5.1.2 Selection of explanatory variables

To select the explanatory variables to be in the logistic model, we can use stepwise method. Stepwise variable selection algorithms as in ordinary regression, algorithms can select or delete predictors from a model in a stepwise manner.

In exploratory studies, such model selection methods can be informative if used cautiously, in the method we can use two approaches: forward selection, or backward selection. With either approach, the process should consider the entire variable at any stage rather than just individual indicator variables, otherwise, the result depends on how you choose the baseline category for the indicator variables, add or drop the entire variable rather than just one of its indicators. In any case, statistical significance should not be the sole criterion for whether to include a term in a model. It is sensible to include a variable that is important for the purposes of the study and report, its estimated effect even if it is not statistically significant. Likewise, with a large sample size sometimes a term might be statistically significant, but not practically significant, we might then exclude it from the model because the simpler model is easier to interpret, Agresti (2007).

3.5.1.3 Strategies in selection of model

The selection process becomes harder when the number of explanatory variables or predictors increases. There are two competing goals: the model should be simple to interpret the research questions. Most studies are designed to answer certain questions. The questions guide the choice of model terms. Confirmatory analyses use a restricted set of models like a study hypothesis of effects. For studies that are exploratory rather than confirmatory, a search among possible models may provide evidence

about the dependence structure and raise questions for future research. It is helpful first to study the effect on Y of each predictor by itself using graphics. For a continuous or a discrete predictor, this gives a ‘‘feel’’ for the marginal effects. The question that arises now is how many predictors can we use? also the limits of the number of predictors, for which effects can be estimated precisely? Agresti (2007) suggested a guideline that the ideally outcomes at least should be 10 outcomes for each type for every predictor. The problem here, with many predictors sometimes we face type of multicollinearity problem. Correlation among predictors leads to no one variable is important when all others are in the model. Also a variable may has a little effect because it is overlap considerably with other predictors in the model. Deleting such a redundant predictor can be helpful to reduce standard errors of other estimated effects. Agresti (2002), proposed methods analogous to forward selection and backward elimination in ordinary regression. Forward selection adds terms sequentially until further additions do not improve the fit. At each stage it is selecting the term giving the greatest improvement in fit. The minimum P-value for testing the term in the model is a sensible criterion, since reductions in deviance for different terms may have different df values. Stepwise variation of this procedure retests, each stage, terms added at previous stages to see if they are still significant. Backward elimination begins with a complex model and sequentially removes terms, at each stage. It is selected the term for which its removal has the least damaging effect on the model (e.g., largest P-value), the process stops when any further deletion leads to a significantly poorer fit. With either approach, for qualitative predictors have more than two categories. The process should consider the entire variable at any stage rather than just individual dummy variables, add or drop the entire variable rather than just one of its dummies. Many

statisticians prefer backward elimination over forward selection, feeling it is safer to delete terms from an overly complex model than to add terms to an overly simple one. Finally, statistical significance should not be the sole criterion for inclusion of a term in a model. It is sensible to include a variable that is central to the purposes of the study and report its estimated effect even if it is not statistically significant. Keeping the variable in the model may help reduce the bias in estimated effects of other predictors, and it can make it possible to compare results with other studies where the effect is significant.

3.6 The goodness-of-fit of multiple logistic regression model

After selecting a preliminary model, we obtain further insight by switching to a microscopic mode of analysis, such diagnostic analyses may suggest a reason for the lack of fit, such as nonlinearity in the effect of an explanatory variable. Two tests can be used to chick the goodness-of-fit, Pearson chi-square test for goodness-of-fit and likelihood ratio test. Both are chi-square methods, the Pearson statistic is based on traditional chi-square, and the likelihood ratio test statistic is based on likelihood ratio chi-square. The likelihood ratio test is preferred over the Pearson. Either test is preferred over classification tables when assessing model fit. Both tests usually yield the same substantive results, Garson (2009).

3.6.1 Pearson Chi- square test as goodness-of-fit

Karl Pearson developed the goodness-of-fit to determine whether observed distributions of frequency data fitted a theoretical distribution. By estimating the cell frequencies one would expect to observe under some theoretical distribution, one could compare the actual frequencies and calculate the difference. The smaller the difference is the better the fit. Using the test statistic and degree of freedom, we can estimate the

significance value. The assumptions underlie the test is the data must be treated as categorical. It must be mutually exclusive, none of expected values may be less than 1, and no more than 20% of expected values may be less than 5, each observation is independent of each other observation, Salkind (2007).

3.6.2 The Likelihood Ratio Test

Here we will distinguish between the function of the likelihood, log likelihood, log likelihood ratio, and the likelihood ratio test. The likelihood is a probability, specifically the probability that the observed values of the response variable may be predicted from the observed values of the explanatory variables. Like any probability, the likelihood varies from 0 to 1. The function of log likelihood is used in significance testing in multinomial logistic regression analysis. The log likelihood (LL) is its log and varies from 0 to - ∞, as the log of any number less than 1 is negative, (LL) is calculated through iteration, using ML estimation. LL is basis for test of a logistic model, (-2LL) has approximately a chi-square distribution, (-2LL) can be used for assessing the significance of MLR. The (-LL) statistic is the likelihood ratio. Also is called goodness-of-fit, deviance chi-square, and also called scaled deviance or deviation chi-square. The (-2LL) reflects the significance of the unexplained variance in the response variable. In general, as the model becomes better, -2LL will decrease in magnitude. The likelihood ratio is defined in several different but essentially equivalent ways, one is

$$\Lambda = \frac{\sup_{\theta \in \Theta} L(\theta, y)}{\sup_{\theta \in \Theta_{\circ}} L(\theta, y)}$$

the likelihood ratio is not used directly in significance testing, but it is the basis for the likelihood ratio test, which

is the test of the difference between two likelihood ratios of two of -2LL's. The likelihood ratio test is a test of the significance of the difference between the likelihood ratio (-2LL) for the researcher's model minus the likelihood ratio for a reduced model. This difference is called "model chi-square".

The likelihood ratio test is generally preferred over its alternative, the Wald test. The likelihood ratio test looks at model chi-square by subtracting deviance (-2LL) for the final (full) model from deviance for the intercept only model, df in this test equal the number of terms in the model minus 1 (for the constant). This is the same as the difference in the number of terms between the two models, since the null model has only one term. Model chi-square measures the improvement in the explanatory variables make compared to the null model. If the LL test statistic shows a small p-value (<=0.05) for a model with large effect size, ignore contrary finding as it is biased toward type II errors, in such instances instead assume good model fit overall. Garson (2009). A common use of the likelihood ratio test is testing the difference between a full model and reduced model dropping an interaction effect. If model chi-square (which is -2LL for the full model minus -2LL for the reduced model) is significant, the interaction effects is contributing significantly to the full model. Likelihood ratio test assesses the overall logistic model, but does not tell us if particular explanatory variables are more important than others. This can be done by comparing the difference in -2LL for the overall model with a nested model which drops one of the explanatory variables. A non-significance likelihood ratio test indicates no differences between the full and reduced model. The likelihood ratio test of individual parameters is better criterion than alternative Wald statistic when considering which variables to drop from the logistic regression

model, Garson (2009). In general, the likelihood ratio test can be used to test the difference between a given model and any nested model which is a subset of the given model. Likelihood ratio test cannot be used to compare two non-nested models. Chi- square can be used to help decide which variables to drop from or add to the model. If the test is significance we conclude that the variable dropped in the nested model do matter. If the test is not significance we conclude the variable makes no difference from the model.

3.6.2.1 The likelihood-ratio test as goodness -of- fit test

The likelihood-ratio statistic $-2(L_o - L_1)$ tests whether certain model parameters are zero, this can be done by comparing the log likelihood L_1 for the fitted model M_1 with L_o for a simpler model M_o , denote this statistic for testing M_o, given that M_1 holds, by $G^2(M_o|M_1)$. The goodness-of-fit statistic $G^2(M)$ is a special case in which $M_o = M$ and

M_1 is the saturated model, in testing whether M fits, we test whether all parameters in the saturated model but not in M equal zero. The asymptotic df is the difference in the number of parameters in the two models, which is the number of binomials modeled minus the number of parameters in M. Let L_s denotes the maximized log likelihood for the saturated model. The likelihood-ratio statistic for comparing models M_1 and M_o is:

$$G^2(M_o|M_1) = -2(L_o - L_1) - [-2(L_1 - L_s)] = G^2(M_o) - G^2(M_1) \quad (3.16)$$

The test statistic comparing two models is identical to the difference in

G^2 Goodness-of-fit statistics (deviances). For the two models, comparison statistic often has an approximate chi-squared null distribution even when separate $G^2(M_i)$ do not. Nonetheless, if df for the comparison statistic is modest (as in comparing two models that differ by a few parameters), the null distribution of $G^2(M_o|M_1)$, is approximately chi-squared.

3.6.3 Checking of goodness-of-fit

In practice, there is no guarantee that a certain logistic regression model fits the data well. For any type of binary data, one way to detect lack of fit is using a likelihood-ratio test to compare the model with more complex ones. A more complex model might contain a nonlinear effect, such as a quadratic term. Models with multiple predictors would consider interaction. If more complex models do not fit better, this provides some assurance that the model chosen is reasonable. Other approaches to detecting lack of fit are using a Pearson χ^2 or likelihood-ratio G^2 statistic. The saturated model differs in the two cases, an asymptotic chi-squared distribution for the deviance results as n $\rightarrow\infty$ with a fixed number of parameters in that model and hence a fixed number of settings of predictor values.

3.7 Diagnostics of multiple logistic regression model

Diagnostics apply to any test, measurement, or decision-making strategy that categorizes people. It can be examined the relationship between how a test categorizes a subject and in which category the subject actually is, relevant categories might include, among others. Diagnostics usually follow some procedures like: summarized the input data with a simple way. It indicates the size of the data sets, and it indicates the number of

rows of data (records) that the model processed. Also it displays the means of all the variables, the means of the variables are particularly useful for detecting any problems with the data setup. The goodness-of-fit indices AIC and BIC, both criteria indicate superior model performance the closer they to 0, that is, other things equal, given two models with equal log likelihood values. The model with the fewer parameters is better. Provides information about the parameters and the variance–covariance matrix of the estimates for each segment. This information is useful for identifying the parameters that are statistically Significant. We can interpret these parameters in a manner similar to how interpret regression parameters. A significant variable influences the choice probabilities of each alternative, whereas an insignificant variable does not. Displays the members belonging to each segment, and the segment size, this information are useful for determining the characteristics of individuals in each segment for targeting purposes. Estimate the probability of each selecting case, additional useful diagnostic are the hit ratio, and the average choice probability, which denotes the average of the estimated choice probabilities for the choices actually made by the customers. Hit ratios can be computed using the estimation sample, or part of the data set can be set aside for holdout prediction (i.e., these data are not use in estimating the model parameters), and the hit ratio estimated on the holdout sample. Gives the estimated choice probabilities that can also be used to compute estimated choice share for each alternative in each segment. If a holdout sample is used for prediction, then the choice shares are computed on the holdout sample, rather than on the sample used for model estimation, otherwise, the choice shares are computed on the estimation sample. Provides information about the elasticity of impact of each variable on choice shares. Provides some statistics that can help to evaluate how well the

chosen model. The "full-parametric model" compares to naive model that assigns equal probabilities to all alternative, the reported chi-square value is asymptotically distributed as a chi-square distribution with the indicated df. The goodness-of-fit index provides additional information about the performance of the model. Rho- square is similar to R^2 in regression, which corrects for the number of parameter included in the full-parametric model. Mullender (2005).

3.8 Model validation

Multinomial logistic regression analysis requires that the dependent variable be non-metric. Dichotomous, nominal, and ordinal variables satisfy the level of measurement requirement.

Multinomial logistic regression does not make any assumptions of normality, linearity, and homogeneity of variance for the independent variables.

Regarding the sample size the minimum number of cases per independent variable is 10, using a guideline provided by Hosmer and Lemeshow, authors of *Applied Logistic Regression*, and for preferred case-to-variable ratios, it is better to use 20 cases per one independent variable (20 to 1).

There must be no Problem of Multicolinearity which checked by looking to the standard error (SE) for each predictor, SE less than 2 indicate that there is no multicolinearity problem and ES greater than 2 indicate a problem of multicolinearity.

Chapter Four

Data and application of the method

4-1 Preface

4-2 Variables of the study

4-3 Analysis outcomes and discussion

4-4 The fitted models

4.1 Preface

This chapter aims to apply appropriate methodology for achieving the research study objectives. In order to examine the key determinants of the factors affecting the efficiency of the income satisfaction respondents were asked to provide answers to the questionnaire questions to build the required model.

This chapter outlines the data collection and statistical analyses methods that was used in this research study.

This study employed a categorical data as well as quantitative data collection method using the survey approach.

Data analysis for the final conceptual model was performed using the statistical package of social sciences (SPSS) software version 21.

The means, standard deviation, Minimum and Maximum were calculated for the continuous variables and the percentages for the categorical variables. Likelihood Ratio Tests and chi-square were used to fit the models information. Model accuracy and by chance mode accuracy in addition to Receiver Operating Characteristic Curve (ROC) were also calculated to check the postulated models sensitivity. Likelihood Ratio Tests was also calculated to estimate the parameters and check the level of variables significancy in the models.

4-2 Variables of the study

Variables used in the study including both categorical and continuous variables.

Dependent variable: (Income satisfaction) has household income quite enough to enable families' meet their expenditure requirements? (Categorical) classified into three levels;

i. Low income (i.e. mean income under 1000 SDG)

ii. Middle income (i.e. mean income between 1000 and 2000 SDG)
iii. High income (i.e. mean income above 2000 SDG)

Independent variables (which were categorical, continuous and discrete):

i. Household head age. (X_1) (Continuous)
ii. Household size, (X_2) (mean number of family members). Which is discrete variable.
iii. Household level of Expenditure, (X_3) (Scale)
iv. Household number of student at university level, (X_4) (Discrete)
v. Household number of student at school level, (X_5) (Discrete)
vi. Household monthly income. (X_6)
vii. Household head year of education, (X_7)
viii. Occupation, (X_8) (categorical)
ix. Gender of household head, (X_9) (Dichotomous binary)

Where X_1 is household head age, X_2 is household size, X_3 is household level of expenditure, X_4 is household's numbers of students at universities, X_5 is household's numbers of students at schools, X_6 is household monthly income, X_7 is household head years of education, X_8 is the household head type of occupation and X_9 is the household head gender.

4.3 Analysis outcomes and discussion

Table 4.1: Variables descriptive statistics

	HH.Ag	H.Siz	Exp.L	Unv.Studt	Sch.Stud	Mnth.Inc
N	307	307	307	307	307	307
Mean	43.7	7.51	44.93	0.82	3.13	1279
St. D	12.4	2.72	15.18	0.38	2.21	466
Min	22.00	2.00	15.00	.00	.00	600.00
Max	85.00	16.00	100.00	1.00	11.00	3500.00

Source: Researcher by using SPSS Package

From the table (4.1), the total numbers of observation is 307 cases with 6 metric variables, it is clear that the household head mean age of the selected sample was 44 years old which indicate that they are all matured enough to provide accurate information, with standard deviation equal 12.4 years, the minimum value was 22 years and the maximum value was 85 years.

The mean of the household size is almost 8 persons per household with standard deviation equal to 2.7 and a minimum of 2 persons and maximum of 16 persons per household, the household having more than 12 persons almost due to a household head with two spouses.

Mean level of expenditure per household is 44.9 SDG a day with standard deviation 15.18 SDG maximum and minimum 100 SDG, 15 SDG respectively. Only one student at university level per household is observed as a mean, meanwhile there are many household having no student at university level. Concerning the numbers of student at school level, there is 3.13 student per household as a mean, with maximum number equal to 11 student at school since there is other household also have no student at school level. The mean average of monthly income per household is 1279 SDG with standard deviation equal to 466 SDG which indicate that there is a huge variation in the total monthly income between the household, the minimum monthly income is 600 SDG and the maximum is 3500 SDG.

Table 4.2: Description of household head gender

	Frequency	Percent	Cumulative Percent
Male	240	78.2	78.2
Female	67	21.8	100.0
Total	307	100.0	

Source: Researcher by using SPSS Package

Above table (4.2) showed that there was 240 out of 307 of the respondent were male, while the remaining 67 were female.

Table 4.3: Description of household head education level

	Frequency	**Percent**	**Cumulative Percent**	
Basic	110	35.8	35.8	
Secondary School	61	19.9	55.7	
University	20	6.5	62.2	
Post University	2	.7	62.9	
illiteracy	114	37.1	100.0	
Total	**307**	**100.0**		

Source: Researcher by using SPSS Package

Table (4.3) showed that most of the respondents in the selected sample having education level less than secondary school (72.9%) which may probably affected their level of income. Only 22 of them have university degree

Table 4.4: Description of household head occupation

	Frequency	Percent	Cumulative Percent
Employee	40	13.0	13.0
Worker	70	22.8	35.8
Farmer	123	40.1	75.9
Trader	62	20.2	96.1
Other	12	3.9	100.0
Total	**307**	**100.0**	

Source: Researcher by using SPSS Package

Regarding the type of occupation for the household head it is clear that from the table (4.4) agriculture represent the main source of their income, that is even the number of workers appear in the above table were doing their jobs as a workers in farms (66.9%)

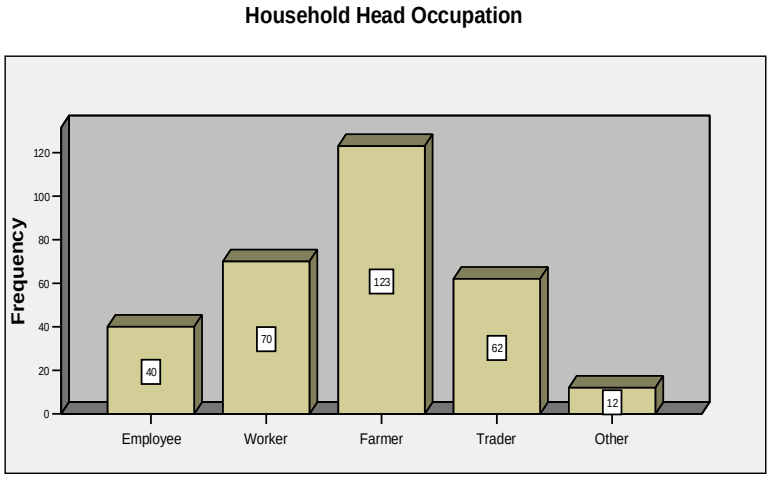
Table 4.5: Description of household income sources

	Frequency	Percent	Cumulative Percent
Agriculture	126	41.0	41.0
Trade	56	18.2	59.3
Livestock	5	1.6	60.9
Labour	69	22.5	83.4
Empolyment	40	13.0	96.4
Other	11	3.6	100.0
Total	307	100.0	

Source: Researcher by using SPSS Package

The table no (4.5) consolidate what we mentioned in the table no (4-4) that the main source of income for the household head is come from the agricultral activities (both agriculture, labour and livestock) which reprent about 67% of their income sources.

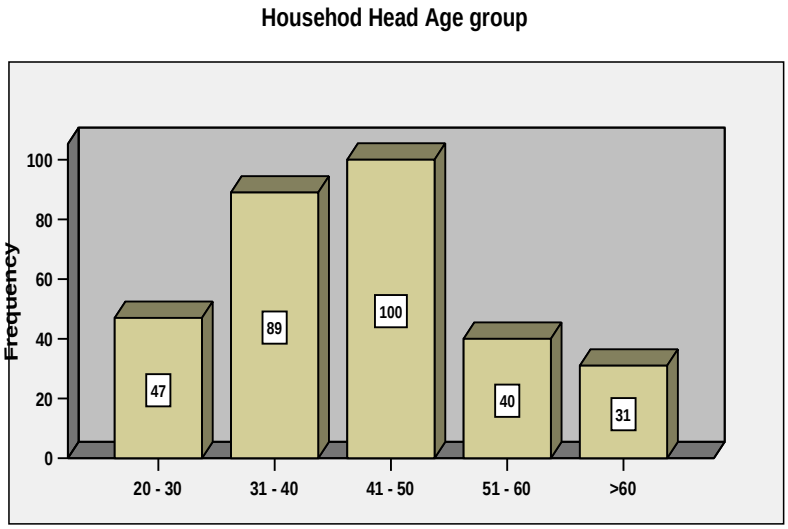
Figure 4.1: Household head occupation



Source: Researcher by using SPSS Package

The figure (4.1) showed that most of the respondent were farmers followed by workers, traders, employee and others.

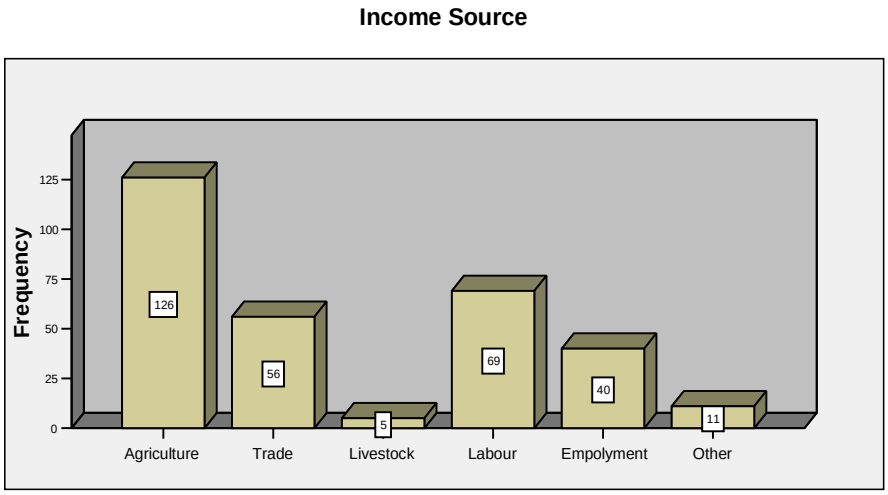
Figure 4.2: household head age group



Source: Researcher by using SPSS Package

Figure no (4.2) showed that almost most of the respondent in the selected sample having an age over than 40 years of old except only about 50 respondents their age is between 20 to 30 years of old.

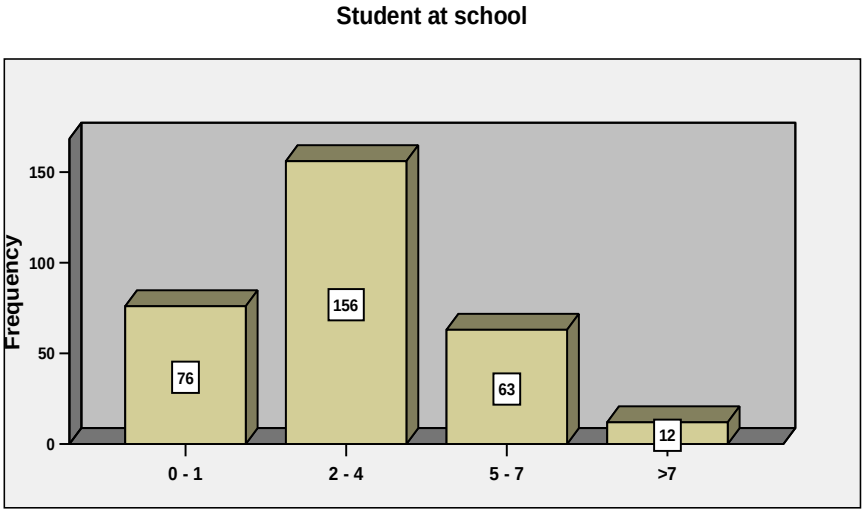
Figure 4.3: Household income source



Source: Researcher by using SPSS Package

Also its clear that from the above figure (4.3) that the agricultural activities remain the main sources of income for the selected respondents.

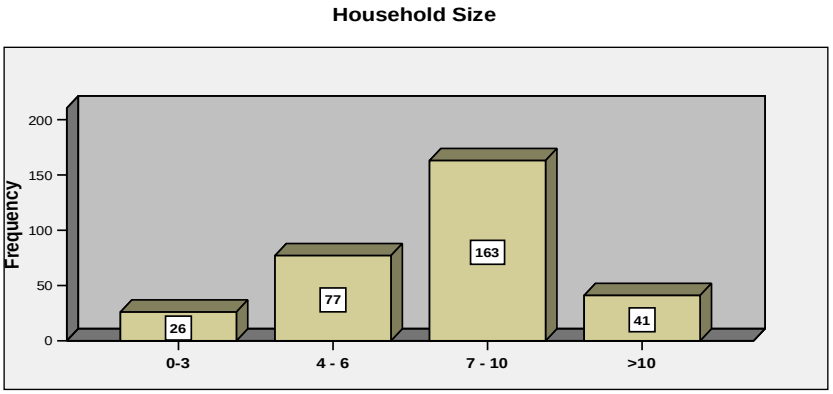
Figure 4.4: Number of student at school level



Source: Researcher by using SPSS Package

Figure no (4.4) showed that about 156 of the household having 2 to 4 student at school level followed by 76 household having 0 to 1 student at school level, 63 of the households having 5 to 7 students at school levels and only 12 households having more than seven students at school level.

Figure 4.5: Household size



Source: Researcher by using SPSS Package

From the above figure (4.5) it is obvious that there is 163 households with an average size ranged between 7 to 10 persons followed by 77 households with an average number of persons ranged between 4 to 6, 41 households with an average size of more than 10 persons and only remaining 26 household with an average persons between 0 and 3.

Figure 4.6: Household head gender

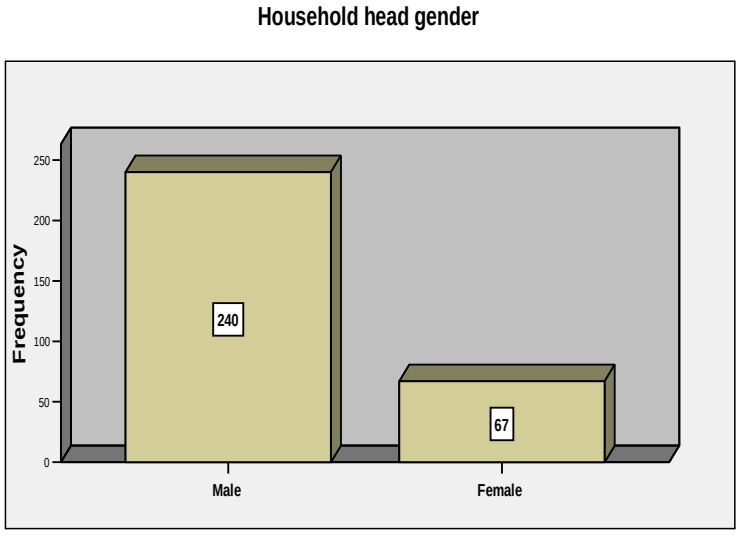


Figure (6), Source: Researcher by using SPSS Package

The figure (4.6) showed that 240 of the respondent appear in the sample of the study were male and only 67 of them were female.

Table 4.6: Frequencies of the response variable categories

Income satisfaction	Frequency	Percent
less satisfy income	70	22.8%
To some extent satisfy income	192	62.5%
Quite satisfy income	45	14.7%
Total	**307**	**100%**

Source: Researcher by using SPSS Package

Table 4.7: Model fitting information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	560.017			
Final	194.748	365.269	18	.000

Source: Researcher by using SPSS Package

Model Fitting Information in the above table (4.7) indicates the parameters of the model for which the model fit is calculated. "Intercept Only" describes a model that does not control for any predictor variables and simply fits an intercept to predict the outcome variable.

The last row in the table "Final" describes a model that includes the specified predictor variables and has been arrived at through an iterative process that maximizes the log likelihood of the outcomes seen in the outcome variable. By including the predictor variables and maximizing the log likelihood of the outcomes seen in the data, the "Final" model should improve upon the "Intercept Only" model. This can be seen in the differences in the -2(Log Likelihood) values associated with the models.

The -2(Log Likelihood) is the product of -2 and the log likelihoods of the null model and fitted "final" model. The likelihood of the model is used to test of whether all predictors' regression coefficients in the model are simultaneously zero and in tests of nested models.

Chi-Square is the Likelihood Ratio (LR) Chi-Square test that at least one of the predictors' regression coefficient is not equal to zero in the model. The LR Chi-Square statistic can be calculated by -2*L(null model) - (-2*L(fitted model)) = 560.017 - 194.748 = 365.269, where L(null model)

is from the log likelihood with just the response variable in the model (Intercept Only) and L(fitted model) is the log likelihood from the final iteration (assuming the model converged) with all the parameters.

df indicates the degrees of freedom of the chi-square distribution used to test the LR Chi-Square statistic and is defined by the number of predictors in the model (nine predictors in two models).

Sig. - This is the probability getting a LR test statistic being as extreme as, or more so, than the observed statistic under the null hypothesis; the null hypothesis is that all of the regression coefficients in the model are equal to zero. In other words, this is the probability of obtaining the chi-square statistic (365.269), or one more extreme, if there is in fact no effect of the predictor variables. The p-value is compared to a specified alpha level, our willingness to accept a type I error, which is typically set at 0.05. The small p-value from the LR test, <0.00001, would lead us to conclude that at least one of the regression coefficients in the model is not equal to zero. The parameter of the chi-square distribution used to test the null hypothesis is defined by the degrees of freedom in the prior column.
The presence of a relationship between the dependent variable and combination of independent variables is based on the statistical significance of the final model chi-square.

In this analysis and in the table (4.7) titled "Model Fitting Information" the probability of the model chi-square (365.269) was 0.000, which is less than to the level of significance of 0.05. The null hypothesis that there was no difference between the model without independent variables and the model with independent variables was rejected. Therefore, the existence of a relationship between the independent variables and the dependent variable was supported.

For the Strength of multinomial logistic regression relationship, while multinomial logistic regression does compute correlation measures to estimate the strength of the relationship (pseudo R square measures), these correlations measures do not really tell us much about the accuracy or errors associated with the model.

A more useful measure to assess the utility of a multinomial logistic regression model is classification accuracy, which compares predicted group membership based on the logistic model to the actual, known group membership, which is the value for the dependent variable.

The benchmark that we will use to characterize a multinomial logistic regression model as useful is a 25% improvement over the rate of accuracy achievable by chance alone.

Even if the independent variables had no relationship to the groups defined by the dependent variable, we would still expect to be correct in our predictions of group membership some percentage of the time. This is referred to as by chance accuracy.

The estimate of by chance accuracy that we will use is the proportional by chance accuracy rate, computed by summing the squared percentage of cases in each group. The only difference between by chance accuracy for binary logistic models and by chance accuracy for multinomial logistic models is the number of groups defined by the dependent variable.

It is clear that from the table (4.8) that the percentage of cases in each group defined by the dependent variable is adopted.

Table 4.8: Case processing summary

		N	Marginal Percentage
Income satisfaction	less satisfy income	70	22.8%
	To some extent satisfy income	192	62.5%
	Quite satisfy income	45	14.7%
Valid		307	100.0%
Missing		0	
Total		307	
Subpopulation		307[a]	

a. The dependent variable has only one value observed in 307 (100.0%) subpopulations.

Source: Researcher by using SPSS Package

Column (N) provides the number of observations fitting the description in the first column. i.e, the first three values give the number of observations for which the category of income satisfaction is being less satisfy income, to some extent satisfy income or quite satisfy income, respectively.

The column of marginal Percentage, lists the proportion of valid observations found in each of the outcome variable's groups. This can be calculated by dividing the N for each group by the N for "Valid". Of the 307 subjects with valid data, 70 are being have less satisfy income in compare with those have to some extent satisfy income and quite satisfy income. Thus, the marginal percentage for this group is (70/307) * 100 = 22.8 %.

In this regression, the outcome variable is income satisfaction (inc.satis) which contains a numeric code divided into three categories. The data includes three levels of inc.satis representing three different groups: 1 = less satisfy income, 2 = to some extent satisfy income and 3 = quite satisfy income.

The second row of the table (i.e, valid) indicates the number of observations in the dataset where the outcome variable and all predictor variables are non-missing.

The third row (i.e, missing) indicates the number of observations in the dataset where data are missing from the outcome variable or any of the predictor variables.

The fourth row (i.e, total) indicates the total number of observations in the dataset--the sum of the number of observations in which data are missing and the number of observations with valid data.

The fifth row (i.e, subpopulation) indicates the number of subpopulations contained in the data. A subpopulation of the data consists of one combination of the predictor variables specified for the model.

The proportional by chance accuracy rate was computed by calculating the proportion of cases for each group based on the number of cases in each group in the 'Case Processing Summary', and then squaring and summing the proportion of cases in each group ($0.228^2 + 0.625^2 + 0.147^2 = 0.464$).

Then the proportional by chance accuracy criteria is equal to 58.02% which is (1.25 x 46.4% = 58.02%).

To characterize our model as useful, we compare the overall percentage accuracy rate in the table no (4.9) with the proportional by chance accuracy.

Table 4.9: Classification

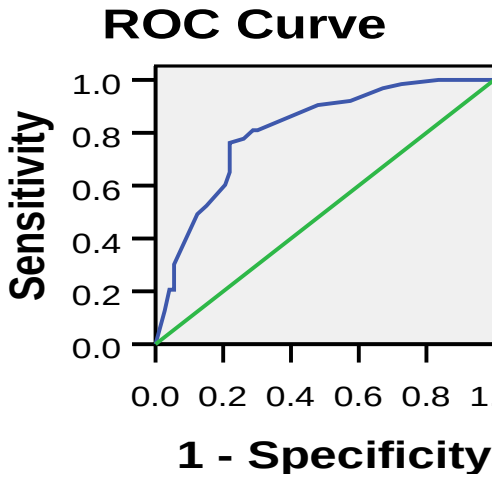
Observed	**Predicted**			
	less satisfy	To some extent	Quite for some extent	Percent Correct
less satisfy income	59	11	0	84.3%
To some	8	180	4	93.8%

extent satisfy income				
Quite enough income	1	3	41	91.1%
Overall Percentage	**22.1%**	**63.2%**	**14.7%**	**91.2%**

Source: Researcher by using SPSS Package

The classification accuracy rate was 91.2% which was greater than or equal to the proportional by chance accuracy criteria of 58.02% (1.25 x 46.4% = 58.02%), so that the criterion for classification accuracy is satisfied in this study.

Figure 4.7: Receiver operating characteristic curve



Source: Researcher by using SPSS Package

To consolidate these results, the receiver operating characteristic curve (ROC) is obtained which is also used to indicate the sensitivity and specificity for all possible cutoff points. The area under the ROC curve ranges from 0.5 and 1.0 with larger values indicative of better fit. The area under the curve is 0.809 with 95% confidence interval (0.737,

0.882). Also, the area under the curve is significantly different from 0.5 since p-value is (0.000) meaning that the logistic regression classifies the group significantly better than by chance.

To see what if the data under the study has a numerical problems or not, we should have to look to the multicollinearity. The multicollinearity in the multinomial logistic regression is detected by examining the standard errors for the b coefficients. A standard error larger than 2.0 indicates numerical problems, Agresti (2002). From the table (4.11) none of the independent variables has a standard error larger than 2.0. (We are not interested in the standard errors associated with the intercept.)

Table No 4.10: Likelihood ratio tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	290.925	90.560	2	.000
hh.age	200.982	.617	2	.734
hh.size	302.335	21.971	2	.038
expn.lev	364.560	23.196	2	.030
univ.studt	202.225	1.861	2	.394
sch.studt	200.773	.409	2	.815
month.inc	452.581	252.216	2	.000
Edu.year	278.128	20.763	2	.041
hh.ocu	203.543	3.179	2	.204
hh.gen	202.015	1.651	2	.438
The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.				

Source: Researcher by using SPSS Package

The table (4.10) shows which of the independent variables is statistically significant. It is obvious that five of the independent variables are not statistically significant because the value of their significance is greater

than the p-value which is equal to 0.05, those are mainly the household head age, number of student at university level, number of student at school level, household head type of occupation and household head gender with value of sig = 0.734, 0.394, 0.815, 0.204 and 0.438 respectively. While the other variables are statistically significant since the values of their sig less than the p-value (i.e. 0.05) those are; the level of expenditure, the total of monthly income and the mean number of the household size with sig value = 0.030, 0.00 and 0.038 respectively (the "Sig." column). There is not usually any interest in the model intercept (i.e., the "Intercept" row).

Baseline category (reference) of the response variable

Any category of the response variable can be chosen to be the baseline or reference

category, the model will fit equally well, achieving the same likelihood and producing the same fitted values, only the values and interpretation of the parameters will change, Schafer (2006). In the current study we choose the quite satisfy income category which means, the comparison will be against the quite satisfy income category.

Table 4.11: Parameter estimates

income satisfaction[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
less satisfy	Intercept	32.016	4.582	48.825	1	.000			
	hh.age	.011	.039	.086	1	.769	1.011	.937	1.092
	hh.size	-.039	.020	3.978	1	.046	.962	.926	0.999
	expn.lev	.817	.391	4.366	1	.037	2.263	1.052	4.869
	univ.studt	-.985	1.024	.926	1	.036	.373	.050	2.777

	sch.studt	-.096	.214	.200	1	.654	.909	.598	1.381
	month.inc	-.025	.003	72.046	1	.000	.975	.969	.981
	Edu.year	.044	.089	.247	1	.619	1.045	.878	1.245
	hh.ocu (w)	-.088	.415	.045	1	.832	.916	.407	2.064
	hh.gen (f)	.335	1.019	.108	1	.742	1.398	.190	10.290
To some extent satisfy	Intercept	16.498	3.516	22.016	1	.000			
	hh.age	.021	.034	.376	1	.540	1.021	.955	1.091
	hh.size	-.147	.162	.823	1	.364	.863	.628	1.186
	expn.lev	-.866	.424	4.171	1	.041	.421	.183	0.966
	univ.studt	.029	.684	.002	1	.966	1.029	.269	3.932
	sch.studt	-.105	.164	.408	1	.523	.900	.652	1.243
	month.inc	-.010	.002	34.390	1	.000	.990	.987	.994
	Edu.year	-1.724	.743	5.387	1	.020	.178	.042	0.765
	hh.ocu (w)	.315	.295	1.140	1	.286	1.371	.768	2.446
	hh.gen (f)	-.423	.778	.296	1	.587	.655	.143	3.009

a. The reference category is: Quite satisfy income.

Source: Researcher by using SPSS Package

The above table (4.11) presents the parameter estimates (also known as the coefficients of the model). As we can see, there is no overall statistical significance value. As the dependent variable classified into three categories, we can see that there are two sets of logistic regression coefficients. The first set of coefficients are found in the category "less satisfy income" row (representing the comparison of the less satisfy income category to the reference category, which is (the quite satisfy income category). The second set of coefficients is found in the category "to some extent satisfy income" row (this time representing the comparison of the to some extent satisfy income category to the reference category, (quite satisfy income).

From the table, to see what are the variables that have significant impact on the dependent variable we have to look to the column of sig and compare it by the p-value which is equal to 0.05. If the significance value

is less than or equal to 0.05 in this case we adopted the rejection of the null hypothesis that (H_0: $B_1 = 0$) and accept (H_1: $B_1 = 0$) or there is a significant relation between the independent variable X_1 and the dependent variable.

Concerning the less satisfy income group it is obvious that from the table (4-11) that there are four explanatory variables have significant relationship with the regressor (dependent) variable, these are;

1. The size of household where the value of sig = (0.046) which is < 0.05.
2. The level of expenditure where the value of sig = (0.037) which is < 0.05
3. The numbers of student at university level where the value of sig = (0.036) which is < 0.05 and
4. The level of monthly income where the value of sig = (0.00) which is also < 0.05

Regarding the second category, “to some extent satisfy income group”, we found that there are only three explanatory variables have significant relationship with the repressor (dependent) variable, these are the;

1. The numbers of household head years of education where the value of sig = (0.020) which is < 0.05.
2. The level of expenditure where the value of sig = (0.041) which is < 0.05 and,
3. The level of monthly income where the value of sig = (0.00) which is also < 0.05

Hence, we can conclude that the number of students at university level being within the household in the less satisfy income group had more effects on the efficiency of their income in compare with the number of students at university level being within the household in to the some extent satisfy income group, which indicate that households within this group (to some extent satisfy income group) were more likely able to manage and control their income efficiency. Also, household head numbers of education year in the less satisfy income group showed no significant effect on the income satisfaction, while it has significant impact on the income satisfaction for to the some extent satisfy income group.

In comparing the size of household within the two groups (less satisfy income group and to the some extent satisfy income group) we clearly see that the household size has different impact on the income satisfaction for the two groups. However, in the less satisfy income group the variable (household size) has significant impact on the dependent variable while it has no significant impact on the dependent variable for the second group (to some extent satisfy income group).

The Bs column is the estimated multinomial logistic regression coefficients for the models. An important feature of the multinomial logit model is that it estimates k-1 models, where k is the number of levels of the outcome variable. In this instance, SPSS (Software used for the analyses) is treating the (quite satisfy income) as the referent group and therefore we have two models to be estimated since there is three levels of the dependent variable, first estimated a model for the less satisfy income group relative to the quite satisfy income group and second estimated a model for to the some extent satisfy income group relative to the quite satisfy income group. Therefore, since the parameter estimates

are relative to the referent group, the standard interpretation of the multinomial logit is that for one unit change in the predictor variable, the logit of outcome relative to the referent group is expected to change by its respective parameter estimate (which is in log-odds units) given the other variables in the model being held constant.

The criteria of Odds ratio explanation

The "exp (b)" column in table (4.11) label for odds ratio of the explanatory variables with the response variable, it is predicted change in odds for a unit increase in the corresponding explanatory variable. Odds ratios less than 1 correspond to decreases and odds ratio more than 1.0 correspond to increases. Odds ratios close to 1.0 indicates that unit changes in that explanatory variable does not affect the response variable.

Interpretation of the first group's coefficients relative to the referent group:

First: less satisfy income relative to quite satisfy income:

Intercept - This is the multinomial logit estimate for the less satisfy income relative to quite satisfy income when the predictor variables in the model are evaluated at zero. For certain predictor with zero for the rest of the other predictors, the logit for household head being in the less satisfy income group relative to be in the quite satisfy income group is 32.016.

Household head age (hh.age) - This is the multinomial logit estimate for a one unit increase in household head age (in year) for being within the less satisfy income group relative to be within the quite satisfy income group given all the other variables in the model are held constant. However, If a

household head were to increase his age by one year, the multinomial log-odds of a household head being within the less satisfy income group relative to be within the quite satisfy income would be expected to increase by 0.011 unit while holding all other variables in the model constant.

Household size (hh.size) - Which is the multinomial logit estimate for a one unit increase in household size (the number of people being lived in one home) for the less satisfy income group relative to the quite satisfy income group given all the other variables in the model are held constant. If a size of household were to increase by one person, the multinomial log-odds of a household being within the less satisfy income group relative to the quite satisfy income group would be expected to decrease by 0.039 unit while holding all other variables in the model constant.

Household level of expenditure (expn.lev) - This is the multinomial logit estimate for a one unit increase in household level of expenditure (in pound SDG) for the less satisfy income group relative to the quite satisfy income given all the other variables in the model are held constant. So that, If a household were to increase it is expenditure by one pound SDG, the multinomial log-odds of a household being within the less satisfy income group relative to be within the quite satisfy income group would be expected to increase by 0.817 unit while holding all other variables in the model constant.

Student at university (univ.studt) - This is the multinomial logit estimate comparing household having student at university to a household doesn't have student at university level for the less satisfy income group relative to the quite satisfy income group given all the other variables in the model are held constant. The multinomial logit for household doesn't

have any student at university relative to household having student at university is .985 unit lower than for being within the less satisfy income group relative to the quite satisfy income group given all the other predictor variables in the model are held constant. In other words, household doesn't have student at university level are more likely to be within the quite satisfy income group than to be within the less satisfy group.

Student at school level (sch.studt, means the numbers of students at school level per household) - That is the multinomial logit estimate for a one unit increase in the number of students at school level per household for the less satisfy income group relative to the quite satisfy income group given all the other variables in the model are held constant. Hence, If the number of students at school level were to increase by one student in a certain household, the multinomial log-odds of a household being within the less satisfy income group relative to the quite satisfy income group would be expected to decrease by 0.096 unit while holding all other variables in the model constant.

Monthly income: concerning the household monthly level of income - the multinomial logit estimate for a one unit increase in the monthly income per household (in pound SDG) for the less satisfy income group relative to the quite satisfy income given all the other variables in the model are held constant. If the household monthly income is to increase by one Sudanese pound, the multinomial log-odds of a household being within the less satisfy income group relative to the quite satisfy income group would be expected to decrease by 0.025 unit while holding all other variables in the model constant.

Household head numbers of years of education (Edu.year) - This is the multinomial logit estimate for a one unit increase (in years) in the household head numbers of years spent on education. Therefore, for the

less satisfy income group relative to the quite satisfy income given all the other variables in the model are held constant. If a household head is to increase his education level by one year, the multinomial log-odds of a household being within the less satisfy income group relative to the quite satisfy income would be expected to increase by 0.044 unit while holding all other variables in the model constant.

Household head type of occupation (hh. ocu):

Worker - This is the multinomial logit estimate comparing a household head to be worker (the main source of his income was his own work) to a household head to be employee (the main source of his income was the basic regular salary from his employer) for being fall within the less satisfy income group relative to be fall within the quite satisfy income group given all the other variables in the model are held constant. The multinomial logit for worker headed household relative to employee is 0. 088 unit lower for being within the less satisfy group than to be within the quite satisfy income group given all other predictor variables in the model are held constant. In other words, workers headed household are more likely relative to employees headed household to be fall within the quite satisfy income group than to be fall within the less satisfy income group.

Household head gender (hh.gen):

female - This is the multinomial logit estimate comparing a household head to be females to a household head to be males for being fall within the less satisfy income group relative to be fall within the quite satisfy income group given all the other variables in the model are held constant. The multinomial logit for females headed household relative to males is 0. 335 unit higher for being within the less satisfy group than to be within the quite satisfy income group given all other predictor variables in the model are held constant. In other words, females headed household are

more likely relative to males headed household to be fall within the less satisfy income group than to be fall within the quite satisfy income group.

Interpretation of the second group's coefficients relative to the referent group:

Second: To some extent satisfy income group relative to the quite satisfy income group:

Intercept - This is the multinomial logit estimate for household head being lies in to some extent satisfy income group relative to lies into quite satisfy income group when the predictor variables in the model are evaluated at zero. For certain predictor with zero for the rest of the other predictors, the logit for household head being lies into to some extent satisfy income group relative to the quite satisfy income is 16.498

Household head age (hh.age) - This is the multinomial logit estimate for a one unit increase in household head age (in year) for to some extent satisfy income group relative to the quite satisfy income given all the other variables in the model are held constant. Hence, If a household head age were to increase by one year, the multinomial log-odds of a household being lies within to some extent satisfy income group relative to be within the quite satisfy income group would be expected to increase by 0.044 unit while holding all the other variables in the model constant.

Household size (hh.size) - Which is the multinomial logit estimate for a one unit increase in household size (the number of people being lived in one home) for to some extent satisfy income group relative to the quite satisfy income group given all the other variables in the model are held constant. However, If a size of certain household were to increase by one person, the multinomial log-odds of a household being lies within the to some extent satisfy income group relative to quite satisfy income group

would be expected to decrease by 0.147 unit while holding all other variables in the model constant.

Household level of expenditure (expn.lev) - This is the multinomial logit estimate for a one unit increase in household level of expenditure (in pound SDG) for to some extent satisfy income group relative to the quite satisfy income group given all the other variables in the model are held constant. However, If a household head were to increase his level of expenditure by one pound SDG, the multinomial log-odds of a household being lies within to some extent satisfy income group relative to the quite satisfy income groups would be expected to decrease by 0.866 unit while holding all other variables in the model constant.

Student at university (univ.studt) - This is the multinomial logit estimate comparing a household having student at university level to a household doesn't have student at university level for to some extent satisfy income group relative to the quite satisfy income group given all the other variables in the model are held constant. The multinomial logit for household doesn't have any student at university level relative to household having student at university level is 0.029 unit higher for being lies within to the some extent satisfy income group relative to the quite satisfy income group given all other predictor variables in the model are held constant. In other words, household doesn't have student at university level are more likely to be within quite satisfy income group than to be lies within to the some extent satisfy income group.

Student at school level (sch.studt), means the numbers of students at school level per household - That is the multinomial logit estimate for a one unit increase in the number of students at school per household for the to some extent satisfy income group relative to the quite satisfy income group given all the other variables in the model are held constant. So that, If the number of students at school were to increase by one

student in a certain household, the multinomial log-odds of a household being lies within to the some extent satisfy income group relative to the quite satisfy income group would be expected to decrease by 0.105 unit while holding all other variables in the model constant.

Monthly income: Regarding the household monthly level of income - the multinomial logit estimate for a one unit increase in the monthly income per household (in pound SDG) for to the some extent satisfy income group relative to the quite satisfy income group given all the other variables in the model are held constant. If a household monthly level of income is to increase by one Sudanese pound, the multinomial log-odds of a household being lies within to the some extent satisfy income group relative to the quite satisfy income group would be expected to decrease by 0.010 unit while holding all other variables in the model constant.

Household head numbers of years of education (Edu.year) - This is the multinomial logit estimate for a one unit increase in the household head numbers of years that he spent on education for to the some extent satisfy income group relative to the quite satisfy income group given all the other variables in the model are held constant. If a household head is to increase his education by one year, the multinomial log-odds of a household being lies within to the some extent satisfy income group relative to the quite satisfy income group would be expected to decrease by 1.724 unit while holding all other variables in the model constant.

Household head type of occupation (hh. ocu):

Worker - This is the multinomial logit estimate comparing a household head to be worker (the main source of his income was his own work) to a household head to be employee (the main source of his income was the basic regular salary from his employer) for being fall within to the some extent satisfy income group relative to be fall within the quite satisfy

income group given all the other variables in the model are held constant. The multinomial logit for worker headed household relative to employee is 0. 315 unit higher for being within to the some extent satisfy income group than to be within the quite satisfy income group given all other predictor variables in the model are held constant. In other words, workers headed household are more likely relative to employees headed household to be fall within to the some extent satisfy income group than to be fall within the quite satisfy income group.

Household head gender (hh.gen):

female - This is the multinomial logit estimate comparing a household head to be females to a household head to be males for being fall within to the some extent satisfy income group relative to be fall within the quite satisfy income group given all the other variables in the model are held constant. The multinomial logit for females headed household relative to males is 0. 423 unit lower for being fall within to the some extent satisfy income group than to be fall within the quite satisfy income group given all other predictor variables in the model are held constant. In other words, males headed household are more likely than females headed household to be fall within to the some extent satisfy income group than to be fall within to the quite satisfy income group.

Hence, from the information in table (4.11) we can fit the two logistic regression models as follow:

4-4 The fitted models:

4.4.1 Models with all variables:

$$\log\left[\frac{\pi_j(x_i)}{1-\pi_k(x_i)}\right] = 32.016 + 0.01x_1 - 0.04x_2 + 0.82x_3 - 0.99x_4 - 0.09x_5$$
$$- 0.03x_6 + 0.04x_7 - 0.09x_8 + 0.34x_9 \rightarrow (4.1)$$

$$\log\left[\frac{\pi_j(x_i)}{1-\pi_k(x_i)}\right] = 16.48 + 0.02x_1 - 0.15x_2 - 0.87x_3 + 0.03x_4 - 0.11x_5$$
$$- 0.01x_6 - 1.72x_7 - 0.32x_8 - 0.42x_9 \rightarrow (4.2)$$

4.4.2 Models with significance variables:

$$\log\left[\frac{\pi_j(x_i)}{1-\pi_k(x_i)}\right] = 32.016 - 0.04x_2 + 0.82x_3 - 0.99x_4 - 0.03x_6 \rightarrow (4.3)$$

$$\log\left[\frac{\pi_j(x_i)}{1-\pi_k(x_i)}\right] = 16.48 - 0.866x_3 - 0.01x_6 - 1.724x_7 \rightarrow (4.4)$$

Where X_1 is household head age, X_2 is household size, X_3 is household level of expenditure, X_4 is household's numbers of students at universities, X_5 is household's numbers of students at schools, X_6 is household monthly income, X_7 is household head years of education, X_8 is the household head type of occupation and X_9 is the household head gender, for the two models.

4.4.3 Estimating of the response probabilities:

The MLR model has an alternative expression in terms of the responses probabilities, that is $\pi_j = \frac{e^{\alpha_j + \beta_j x}}{\sum e^{\alpha_j + \beta_j x}}$ j=1,2,…..J In our model, we will denote the probability of the quite satisfy income category (referent category) by π_o and the estimated probability by $\hat{\pi}_o$, the less satisfy income category by π_1 and the estimated probability by $\hat{\pi}_1$ and the to some

extent satisfy income category by π_2 and the estimated probability by $\hat{\pi}_2$. The response probability satisfying that the $\sum_{j=0}^{2} \pi_j = 1$.

From table (4.11) of parameter estimates we can calculate these probabilities by two steps:

First, we can calculate the $\log\left[\frac{\hat{\pi}_1}{\hat{\pi}_o}\right]$ and $\log\left[\frac{\hat{\pi}_2}{\hat{\pi}_o}\right]$ as the response variable has three categories, J = 3, which mean that we have two equations as following:

Let $Y_1 = \log\left[\frac{\hat{\pi}_1}{\hat{\pi}_o}\right]$ and $Y_2 = \log\left[\frac{\hat{\pi}_2}{\hat{\pi}_o}\right]$ then,

$$Y_1 = 32.016 - 0.04x_2 + 0.82x_3 - 0.99x_4 - 0.03x_6 \quad (4.5)$$

$$Y_2 = 16.48 - 0.866x_3 - 0.010x_6 - 1.724x_7 \quad (4.6)$$

We cannot make corresponding statement about the in significant variables. As stated by Agresti (2007). It is sensible to include the variables that is important for the purposes of the study.

Second we can calculate $\hat{\pi}_o$, $\hat{\pi}_1$ and $\hat{\pi}_2$ as following:

$$\hat{\pi}_1 = \frac{\exp(y_1)}{1 + \exp(y_1) + \exp(y_2)} \quad (4.7)$$

$$\hat{\pi}_2 = \frac{\exp(y_2)}{1 + \exp(y_1) + \exp(y_2)} \quad (4.8)$$

$$\hat{\pi}_o = \frac{1}{1 + \exp(y_1) + \exp(y_2)} \quad (4.9)$$

Where the exp term or (e) is = 2.71828, which is the base of the system of the natural logarithms.

4.4.4 Predications using multinomial logistic regression model:

Each case consists of a combination of explanatory variables. Prediction is based on classifying this combination in one of the three groups of the response variable. The model estimates the probabilities of this combination of the three groups of the response variable and then according to the largest probability we will classify the case (we have 307 subpopulation in the model). For the application of the model we had selected randomly of two cases of the data and we used the models to predict from which groups we will classify it, the model consists of two equations, to estimate the three response probabilities π_o , π_1 and π_2 we will use equations (4.4) and (4.5).

Table 4.12: Selected cases information

Explanatory variables	Case No (112)	Case No (217)
Household size (X_2)	11	15
Household level of expenditure (X_3)	30	30
Student at university level (X_4)	1	3
Household monthly income (X_6)	1800	1700
Household year of education (X_7)	0	1

Source: researcher

For the case No (112) we can calculate the probability as follows:

$$Y_1 = \log\left[\frac{\hat{\pi}_1}{\hat{\pi}_o}\right] = 32.016 - 0.04x_2 + 0.82x_3 - 0.99x_4 - 0.03x_6$$

$$= 32.016 - 0.04(11) + 0.82(30) - 0.99(1) - 0.03(1800) = 1.186$$

$$Y_2 = \log\left[\frac{\hat{\pi}_2}{\hat{\pi}_o}\right] = 16.48 - 0.866x_3 - 0.010x_6 - 1.724x_7$$

$= 16.48 – 0.866(30) – 0.010(1800) – 1.724(0) = - 27.5$

However, by using equations number (4.7), (4.8) and (4.9) we can calculate the estimated probabilities to occur in each category as follows:

$$\hat{\pi}_1 = \frac{\exp(y_1)}{1 + \exp(y_1) + \exp(y_2)} =$$

$$\frac{\exp(1.186)}{1 + \exp(1.186) + \exp(-27.5)} = 0.766$$

$$\hat{\pi}_2 = \frac{\exp(y_2)}{1 + \exp(y_1) + \exp(y_2)} =$$

$$\frac{\exp(-27.5)}{1 + \exp(1.186) + \exp(-27.5)} = 0.233$$

$$\hat{\pi}_o = \frac{1}{1 + \exp(y_1) + \exp(y_2)} =$$

$$\frac{1}{1 + \exp(1.186) + \exp(-27.5)} = 0.001$$

These probabilities indicated that the case number 112 (random selected household) has probability equal to 0.766 for a household to be within the less satisfy income category, and probability equal to 0.233 to be within the to some extent satisfy income category, and only 0.001of a household to be within the quite satisfy income category. Therefore, we can conclude that a household to be within the less satisfy income category has the largest probability (0.766) in comparing to the other categories.

For the case number (217) also we can calculate the estimated probabilities as follows:

$$Y_1 = \log\left[\frac{\hat{\pi}_1}{\hat{\pi}_o}\right] = 32.016 - 0.04x_2 + 0.82x_3 - 0.99x_4 - 0.03x_6$$

$=32.016 – 0.04(15) + 0.82(30) – 0.99(3) – 0.03(1700) = 2.046$

$$Y_2 = \log\left[\frac{\hat{\pi}_2}{\hat{\pi}_o}\right] = 16.48 - 0.866x_3 - 0.010x_6 - 1.724x_7$$

$= 16.48 – 0.866(30) – 0.010(1700) – 1.724(1) = - 28.224$

However, again by using equations number (4.7), (4.8) and (4.9) we can calculate the estimated probability to occur in each category as follows:

$$\hat{\pi}_1 = \frac{\exp(y_1)}{1+\exp(y_1)+\exp(y_2)} =$$

$$\frac{\exp(2.046)}{1+\exp(2.046)+\exp(-28.224)} = 0.880$$

$$\hat{\pi}_2 = \frac{\exp(y_2)}{1+\exp(y_1)+\exp(y_2)} =$$

$$\frac{\exp(-28.224)}{1+\exp(2.046)+\exp(-28.224)} = 0.01$$

$$\hat{\pi}_o = \frac{1}{1+\exp(y_1)+\exp(y_2)} =$$

$$\frac{1}{1+\exp(2.046)+\exp(-28.224)} = 0.11$$

However, from the calculated probabilities we can also conclude that the case number 217 (random selected household) has probability equal to 0.88 for a household to be within the less satisfy income category, and probability equal to 0.01 to be within the to some extent satisfy income category, and 0.11of a household to be within the quite satisfy income category. Therefore, we can conclude that a household to be within the less satisfy income category has also the largest probability (0.88) in comparing to the other categories.

Chapter Five

Conclusions and recommendations

5-1 Results

5-2 Recommendations

5.1 Results

From the application of the method of multinomial logistic regression on the collected data to find out the main factors that have real impact on the efficiency of the household income in Sudanese household targeting south Darfur state as a case of study during the period of time between 2013 to 2016, the study come out by the following results followed by a numbers of recommendations.

1. The tests that had been carried out in the analysis showed that the model is well fitted the data.

2. The likelihood ratio test with chi-square = (365.269) indicated that the existence of a relationship between the independent variables and the dependent variable was supported.

3. Nine variables have been tested to see if they have significant impact on the household income satisfaction as dependent variable or not. These variables are the household size, the household head year of education, the household level of Expenditure, the household head age, the household head type of occupation, the household head gender, the household number of student at school level, the household number of student at university level and the Household monthly income.

4. There is clear significant relationship between the independent variables size of household, level of monthly income, level of expenditure and the number of student at the university level for the first income group, while the other remaining independent variables in the same group showed no significant relationship to the dependent variable, these are, the household head year of education, the household head age, the household head type of occupation, the household head gender and the household number of student at school level.

5. For the second group, only three independent variables have significant relationship to the dependent variable, and these are; the level of expenditure, the level of monthly income and the numbers of household head years of education, while the remaining

independent variables have not significant relationship to the dependent variables, these are; the household head age, the household head type of occupation, the household head gender, the household number of student at school level, the household number of student at university level and the household size.

6. The most significant explanatory variables within the two groups having more impact on the dependent variable were the level of income, level of education of the household head and the number of students at university level respectively.

7. The model ability of prediction had been checked through the calculation of the probability of a two cases of the observation selected at random bases and applying the model to predict in which category of the response variable does the case is classified, however, the household is more likely to be within the less satisfy income category with a probability equal to 0.76 and 0.88 for the two selected cases.

8. The logistic regression model is a suitable model to many types of data when the response variable is categorical with more than two categories. Multinomial logistic regression has no any restrictions about the explanatory variables as well as the statistical assumptions; this model is most common in the categorical data analysis. MLR can be used in many areas of social, economics, educational, health, behavioral and even scientific experiments.

5.2 Recommendations

1. Reviewing of the policies and strategies pursued by local authorities relating to prices, and re- evaluate the markets conditions on regular basis and follow up traders and producers at all levels.

2. Low income and limited job opportunities reduce households' capacity in acquiring of the basic needs. However, improved access

to income generating activities is key factor to raise the level of income for the households.

3. Establishing of workshops, conferences and seminars in order to come up with proposals and clear views that help to reduce basic goods prices in order to stop one of the biggest challenges standing in front of achieving social welfare for citizens.

4. Focusing on high level education and makes it available for all groups of people especially vulnerable social groups because of its important role in raising the per capita income level

5. Designing and formulating of more appropriate studies and researches, to propose solutions and educate individuals turn as one of the most important market forces that should put pressure on goods producers and sellers to force them slow down and reduce the commodities prices.

6. Local authorities have to focus on the development of the agricultural sector and should take advantage of the available features concerning the agriculture resources so as to increase agricultural products to the extent that makes it affordable for everyone in the society especially those with low level of income.

7. Encouraging and supporting industrial sector in order to makes it able to provide and secure suitable amount of local commodities at reasonable prices for all people.

8. Adopting suitable mechanisms to reduce the burden of tuition costs, educational fees and Exempting people of low-income groups

9. Establishing of new roads and rehabilitate the old ones in a way that would makes it easy to transfer the production from its areas of origin to the consuming areas (this will probably reduce production cost and guaranteed safe access for the needy people)

10. Taking advantage of the geographical unique location of the state and the preferential advantages in terms of the availability of natural resources through the creation of a favorable and attractive environment of investment to attract foreign investors to invest in the fields of agriculture and complementary industries, through which we can expand and increase the size of consuming production and in the long turn leads to reduce the production prices.

11. Achieving the above recommendation (No 7) will help in the creation of new employment opportunities for the people having no jobs then contribute in reducing the high level of unemployment rate; hence leads to increase the level of personal income of the people and thereby increase their ability to get basic goods and services in an easy way.

12. Creating new mechanisms not only limited to the oversight of the markets but also to support and control the prices through offering goods at affordable prices for all groups of people especially low-income groups.

13. Reducing the amount of goods imported from other countries since there was great deterioration in the local currency against the foreign currency.

14. Oversight tools for pricing should be activated together with giving high concentration on formulating marketing policies including electronic database affording goods and prices information for consumers in a way that enable them easily reach to the reasonable commodities prices.

15. Getting use of the modern technology by adopting the idea of the electronic markets which reduces the costs of production and reflected in minimizing the level of goods prices.

16. Planning for the work of a unified economic system will help to achieve self-sufficiency in essential goods and services and the launch of political freedoms with strict application of the law in respect of monopoly and price manipulation

17. Government should provide workers with sufficient earnings to provide for a family

18. Adopting of specific policy and strategies including strengthening the bargaining power of workers, expanding the earn income, and increasing the minimum wage.

19. Helping workers get the training and education they need to succeed then can expanding their opportunities for getting work with high wages.

20. Household heads have to take into consideration suitable size of their families.

21. Household heads have to raise the level of their education

22. Household heads have to look for additional sources of income

References

1. Abdalla, M. (2012), “An Application on Multinomial Logistic Regression Model” Pak.j.stat.oper.res. Vol.VIII No.2 2012 pp 271-291
2. Agresti. A, (1990), “Categorical data analysis” John Wiley & Sons, Inc. New York. 3rd edition
3. Agresti, A. (2002), "Categorical Data Analysis", John Wiley & Sons, Inc. New York. http://dx.doi.org/10.1002/0471249688.
4. Agresti, A. (2002) “Categorical data analyses” John Wiley & Sons, Inc. New York. second edition
5. Agresti, A. (2007), "An Introduction to Categorical Data Analysis", John Wiley & Sons, Inc. http://dx.doi.org/10.1002/0470114754.
6. Bayaga, A. (2010), “Multinomial Logistic Regression: Usage and Application in Risk Analysis” Journal of Applied Quantitative Methods, Vol. 5, No (2), (summer, 2010) pp. 288-296
7. Brannon, D., Barry, T., Kemper, P., Schreiner, A., and Vasey, J. (2007). Job Perception and Intent to Leave Among Direct Care Workers: Evidence from the Better Jobs Better Care Demonstrations. The Gerontologist 47:820-829, The Geronotological Society of America.
8. Hombres, B Anke Weber and Leandro Elia (2012), "Literature review on income inequality and the effects on social outcomes", JRC scientific and policy reports. http://www.jrc.ec.europa.eu/
9. Liberda, B. and Peczkowski, M. and Ewa Gucwa-Lesny University of Warsaw. “How do we value our income from which we save”
10. Chinhui Juhn, Kevin M. Murphy and Pierce, B. “Wage Inequality and the Rise in Returns to Skill” Journal of Political Economy, 1993, vol. 101, issue 3, pages 410-42.

11. Checchi D. (2003). Inequality in Incomes and Access to Education. A Cross-country Analysis (1960-95). Labour, 17(2), 153-201.http://dx.doi.org/10.1111/1467-9914.00226.
12. Checchi, D. (2006), “The Economics of Education, Human Capital, Family Background and Inequality”. http://dx.doi.org/10.1037/h0041412.
13. Cochran, W. (1947), Some Consequences When Assumptions for the Analysis of Variances are not Satisfied, Article in Biometrics 3(1):22-38.
14. Chia (2008, p.233) finds that family income constraints do matter in determining whether children participate regularly in sporting activities.
15. Chatterjee, S., and Hadi, A. (2006). Regression Analysis by Example. John Wiley & Sons. http://dx.doi.org/10.1002/0470055464.
16. Dayal Talukder, (2014), “Assessing Determinants of Income of Rural Households in Bangladesh: A Regression Analysis” Journal of Applied Economics and Business Research, JAEBR. 4(2): p 80 - 106
17. Bradley, F. (1987) "Double exponential families and their use in generalized linear regression." Journal of the American Statistical Association 81.395 (1986): 709-721.
18. Filmer and Pritchett (1999), "The Effect of Household Wealth on Educational Attainment", Evidence from 35 countries. Population and development review: 85-120 http://dx.doi.org/10.1111/j.1728-4457.1999.00085.x.
19. Gregorio, Jose and Lee, Jong. (2002) “Education and Income Inequality”: New Evidence from Cross-Country Data. Review of

Income and Wealth, Vol. 48, pp. 395-416, 2002. Available at SSRN: http://ssrn.com/abstract=325165

20. Garson (2009), “Testing Statistical Assumptions”
21. Garson, D. (2009). Logistic Regression with SPSS. North Carolina State University, Public administration Program.
22. Gary S. Becker, (1975) “Human Capital, A Theoretical and Empirical Analysis” Volume Publisher: NBER, Volume ISBN: 0-226-04109-3, (p. 13 - 44), http://www.nber.org/books/beck75-1
23. Galor, O. and J. Zeira (1993), “Income Distribution and Macroeconomics,” Review of Economic Studies, 60: 35 −52.
24. http://en.wikipedia.org/wiki/Logit
25. Joseph Berkson , (1944) Applications of the logistic model to bioassay
26. James K. Lindsey. (1997) “Applying of Generalized linear Models”.
27. Kang, H. Park and Peter, M. Kerr (2008) “Determinants of Academic Performance: A multinomial Logit Approach” Journal of Economic Education, Vo. 21, No, 2 (spring, 1990), pp 101-111
28. Lantz PM, House JS, Lepkowski JM, Williams DR, Mero RP, Chen J. (1998) “Socioeconomic factors, health behaviors, and mortality”: Results from a nationally representative prospective study of US adults. JAMA. 1998; 279(21):1703-1708.
29. M. T. Parvin1 & M. Akteruzzaman (2012), “Factors Affecting Farm and non-Farm Income of Haor Inhabitants of Bangladesh” Progress. Agric. 23(1 & 2): 143 – 150, ISSN 1017-8139
30. Mala, A. (2010) “Multinomial logistic regression model to assess the levels in trans, trans-muconic acid and inferential-risk age group among benzene-exposed group” Indian Journal of Occupational and Environmental Medicine. 14(2): 39–41.

31. Moorman, Sara M. and Deborah Carr. (2008). “Spouses’ Effectiveness as End-of-Life Health Care Surrogates: Accuracy, Uncertainty, and Errors of Overtreatment or under treatment.” The Gerontologist 48(6): 811-9.
32. Nelder and Wedderburn (1972). “Introduction to Generalized Linear model” p 543-544
33. Okurut, F. N., Kagiso, M., Ama, N. O., & Okurut, M. L. (2014). “The Impact of Microfinance on Household Welfare in Botswana”. Botswana Journal of Economics, 12(1), 45-58.
34. Perotti (1993). “Income distribution, political instability and investment”. NBER working paper # 4486
35. Richard D. McKelveya & Zavoina, W. (1975) “A statistical model for the analysis of ordinal level dependent variables”. The Journal of Mathematical Sociology. Volume 4, Issue 1, pages 103-120
36. Riggs (2008), “Using Multinomial Logistic Regression Analysis to Develop a Model of Australian Gay and Heterosexual Sperm Donors’ Motivations and Beliefs” International Journal of Emerging Technologies and Society, Vol. 6, No. 2, 2008, pp: 106 - 123
37. Smith, k. (2007), “Determinants of Soviet Household Income” The European Journal of Comparative Economics. Vol. 4, n. 1, pp. 3-24. ISSN 1824-2979
38. Schwab, (2002). “Sample size guidelines for multinomial logistic regression”
39. Schwab, J. (2007). “Multinomial Logistic Regression Basic Relationships”.www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/Analyzi.

40. Topel, Robert (1994a). “Regional Labor markets and the Determinants of Wage Inequality”. American Economic Review; V.84-#2, pp. 17-22.
41. Takagi, E., Silverstein, M., and Crimmins, E. (2007). IntergenerationalCoresidence of Older Adults in Japan: Conditions for Cultural Plasticity.The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 62:S330-S339, The Gerontological Society of America.
42. Train, K. E. (2009). “Discrete choice methods with simulation”. Cambridge university press.

43.الطيب زين العابدين و محمد الأمين خليفة و منزول عسل و آخرون (2013). "دارفور حصاد الأزمة بعد عقد من الزمان". مركز الجزيرة للدراسات

Appendices

A questionnaire submitted for the purpose of completion of study for a Ph.D. degree

To be completed by Household head or relevant person

Questionnaire No: 0 | | |

Please check (√) or write down correct answer

Income Satisfaction	< 1000 SDG	1000-2000	> 2000 SDG				
Household head	Gender	Male	Female				
Household head	Age	Year					
Household head	Education Level	Basic	Secondary	University		Post University	Non
Household head	Occupation	Employee	Worker	Farmer		Business	Other
Household head	Size of own land	1-5 feddan	6-10 feddan	>10 feddan		Non	
Household	Income Sources	Farming	Trade	Livestock	Labour	Employment	Others
Household	House type	Owned	Rented				
Household	Size of own land	# feddan					
Household	Family Size	#					
Household	Labour force	#					
Household	Expenditure Level	#					
Household	university student	Yes	No				
Household	Student at school	#					
Household	Monthly	#					

	Income		

With regards.

Researcher

South Darfur state population distributed by localities 2014-2015

No	Locality Name	# of Population
1	Baladiat Nyala	344.025
2	Nyala Shamal	265.475
3	Ed Alfursan	245.361
4	Buram	171.176
5	Rehid alberdi	230.609
6	Tulus	313.426
7	Kass	232.334
8	Shataia	53.168
9	Katela	132.397
10	Kabom	212.394
11	Alsalam	107.135
12	Um Dafoug	68.071
13	Sharg Algabal	13.582
14	Netega	115.823
15	Beliel	97.289
16	Mershing	48.337
17	Alwohda	58.001
18	Alradom	153.048
19	Alsunta	150.854
20	Gerieda	103.998
21	Demso	241.942
Grand Total		**3.358.877**

Source**:** South Darfur state, Ministry of Education, general department of educational planning, division of statistics and information, annual statistics book 2014-2015

Sampling structure of the study

No	Cluster	Locality
1	I	Kass
2		Gerieda
3		Buram
4		Ed Alfursan
5		Nyal Wasat
6		Nyala Shamal
7		Beliel
8	II	Alsalam
9		Netega
10		Kabom
11		Sharg Algabal
12		Mershing
13		Demso
14		Alsunta
15	III	Shataia
16		Alradom
17		Um Dafoug
18		Alwohda
19		Katela
20		Tulus
21		Rehid alberdi

Source: researcher

Analysis of the various livelihood activities reported by the households and their contribution to total income resulted in 10 distinct livelihood groups:

No	Main source of income	%
1	Cereals and other crops sales of households	25
2	Sale of food aid, wheel barrow, rickshaw.	2
3	Remittances and kiosk	6
4	Agricultural wage labour, bricks, construction, porter, selling water	5
5	Donkey cart, gifts, tea/food selling, handicrafts	9
6	Salaried work	7%
7	Agricultural Wage labour	17
8	Selling firewood, grass and charcoal	10
9	Selling livestock and products	10
10	Skilled labour and other petty trade	9
Total		**100%**

Source: WFP, State Ministry of Agriculture, State Ministry of Health, N. Darfur, Comprehensive food security assessment. 2011