International Journal of Mathematics and Physical Sciences Research ISSN 2348-5736 (Online)
Vol. 5, Issue 1, pp: (9-23), Month: April - September 2017, Available at: www.researchpublish.com

Measuring the Level of Association between Various Differentials and Household Agricultural Income in Rwanda (2013-2014) using Ordered Logistic Regression Model

Jean-Claude Nzabonimpa[1], Dr. Joseph K. Mung'atu[2], Dr. Marcel Ndengo[3]

[1] Students at Jomo Kenyata University of Agriculture and Technology (JKUAT/KIGALI CAMPUS), Master of Science in applied statistics
[2] Lecturer at Jomo Kenyata University of Agriculture and Technology/Nairobi, Kenya
[3] Lecturer at Jomo Kenyata University of Agriculture and Technology/ Kigali, Rwanda

Abstract: **The purpose of this research was mainly to measure the level of association between various differentials and household agricultural income in Rwanda. Basing on the nature of the variables in this study, an ordered logistic regression model (OLR) was used to statistically measure the relationship between an ordinal dependent variable (Agricultural income) and a set of independent variables. In analysis the dependent variable was classified into 3 groups namely: Low, Medium and High agricultural income. The discussed household agricultural income in this study were mainly generated from the sales of crops, processed products, livestock and livestock products, auto-consumption, rent of land and agricultural equipments. The Fisher test was used to analyze the goodness of fit of the model and hence it indicated that the model was statistically significant as a whole, meaning that the observed data perfectly corresponded to the fitted model. The findings from this study revealed that households (HHs) owning at least one cow have more chances of being either in the middle or in the highest categories of agricultural income. The results also reiterated the fact that HH size has a significant influence on agricultural income of the households, since HHs with large size were more likely to be found in the lower category of agricultural income. Additionally, the data highlighted that the likelihood for a HH to be either in the middle or in the highest categories of agricultural income rises when that household resides in a rural area.**

Keywords: **Household Agricultural income Computation, F-test, Treatment of outliers, Imputation, OLR Model.**

1. INTRODUCTION

1.1 Statement of the problem:

The agricultural sector continues to play a crucial role for development especially in low-income countries where the sector is large both in terms of aggregate income and total labor force [6]. In Schultz's view, agriculture is important for economic growth in the sense that it guarantees subsistence for society, without which growth is not possible. Agriculture is seen as an active sector in the economy. In addition to providing labor and food supply, agriculture plays an active role in economic growth through production and consumption linkages. Recent empirical literature considers that the effect of agricultural progress on poverty alleviation is highly positive [14]. It is not economic growth in general that reduces poverty in developing countries, but the direct and indirect effects of growth in agriculture [18]. Similarly, it is estimated that 1 percent per capita agricultural growth reduces poverty 1.6 times more than the same growth in industry and three times more than growth in the service sector [5]. Farming is Africa's main livelihood: more than two-thirds of Africans

depend on agriculture for their incomes and earnings from agriculture make up around 30% of GDP across the continent. According to Food and Agriculture Organization (FAO) analysis, growth in the agriculture sector is 11 times as effective at reducing poverty as growth in other sectors in sub-Sahara Africa. According to the World Bank, Sub-Saharan African agriculture could, and should, be flourishing and has the right conditions to feed itself: enough fertile farmland, enough water and enough favourable climates. As reported by the International Fund for Agricultural Development (IFAD), the Africa Progress Panel and others, Africa has the potential not only to feed itself, but also to become a major food supplier for the rest of the world. For the case of Rwanda, agriculture is the backbone of country's economy and the majority of households are engaged in some sort of crop or livestock production activity. The agriculture sector is therefore widely regarded as the major catalyst for growth and poverty reduction. The aggregate income in Rwanda is mainly derived from agriculture which is actually almost half of all income whereas about a quarter comes from salaried labour, i.e. wages income [25]. Although the country's Economic Development and Poverty Reduction Strategy (EDPRS) defines a large number of programs to boost the agriculture sector, yet information on agriculture income is still limited and additional further and an in-depth statistical analysis in this area is quasi inexistent. While accelerated growth in agriculture as a whole may be the most promising poverty-reduction strategy currently available to Rwanda. Such strategy needs to be guided by a good understanding of the role of various differentials such as residential areas (urban/rural), provinces and characteristics of household (household size and cow ownership) and how they affect the agricultural income. That is why, this study seeks to provide additional analysis on how various differentials impact or influence the agricultural income of households in Rwanda.

1.2 Objectives:

1.2.1 General Objective:

The main objective of this study was mainly to measure the level of association between various differentials and household agricultural income in Rwanda.

1.2.2 Specific Objectives:

The specific objectives of this study were the following:

1. To determine the association between HHs living in urban and rural setting and their agricultural income levels.
2. To analyze the relationship between households agricultural income levels and their residential province.
3. To determine the association between the household agricultural income levels and size of the household.
4. To analyze the relationship between the household agricultural income levels and cow ownership.

1.3 Research Questions:

1. What is the degree of association between HHs living in urban and rural setting and their agricultural income levels?
2. Is there any relationship between household agricultural income levels and their residential province?
3. What is the degree of association between the household agricultural income levels and size of the household?
4. Is there any relationship between the household agricultural income levels and household cow ownership?

1.4 Justification of the study:

This research was designed with the purpose of obtaining a Master's Degree in Applied Statistics and providing analysis that will contribute to the understanding of the nature of agriculture in Rwanda. It will be also supportive to various sectors namely: Policy makers, institutions and other researchers. This study will also inform policy makers about the priority areas of intervention and help them to know where the Government should allocate more supports to the poor households.

1.5 Scope of the study:

The household agricultural income variable used in this research was computed from secondary data collected by NISR in 2013/14 through EICV4 survey. The base population for this study was all households cultivating any land for crop production to eat or sell, or raising animals over the last 12 months preceding the survey.

2. METHODOLOGY

2.1 Sampling frame:

A *sampling frame* is a complete list of all sampling units that entirely cover the target population. The existence of a sampling frame allows a probability selection of sampling units. For a multi-stage survey, a sampling frame should exist for each stage of selection. In probability sampling, the probability of any element appearing in the sample must be known. Therefore, for this to be accomplished some list should be available from which the sample can be selected. Such a list is called a *sampling frame* and should have the property that every element in the population has some chance of being selected in the sample [19]. As we used the secondary data, the sampling frame for the EICV4 was based on the database of all villages from the 2012 Rwanda Population and Housing Census (2012 RPHC). The primary sampling units (PSUs) were the census enumeration areas (EAs), which were small operational areas defined for the census enumeration. The sample size used for that survey was: 1320 clusters (villages) for the first stage and 14,419 households for the second stage.

2.2 Sampling techniques:

This section discusses briefly on sampling techniques and types of sampling. Information on characteristics of the populations is constantly needed by researchers, public officials, health and social services, and others. For that reason, studies need to be carried out to obtain that information. Due to timeliness, cost and workload almost invariably lead to a selection of respondents. The selected respondents should be as representative of the total population as possible in order to make suitable extrapolations concerning the entire population. The selected respondents constitutes a *sample* and the selection process is called *sampling technique*. Actually, there are two types of samples in sampling: probability (random) sampling and non probability sampling. *A probability sampling scheme* is the one in which every unity in the population has a chance of being selected in the sample. Any sampling method where the probability of selection for some elements of the population can't be accurately determined is named a *non-probability sampling*. It involves the selection of elements based on assumptions regarding the population of interest [29]. For this study, we do not talk much more on non-probability sampling instead, we put much emphases on probability sampling specifically to stratified sampling and systematic sampling techniques since they were the ones used in designing EICV4 survey from which the used data were drown.

2.2.1 Stratification:

Stratification is the process by which the survey population is divided into subgroups or *strata* that are as homogeneous as possible using certain criteria. The purpose of stratification is to enhance the sample representativeness with a given total sample size, thereby reducing sampling errors. In a stratified sample, the sampling error depends on the population variance existing *within the strata* but not *between the strata*. For this reason, it pays to create strata with high homogeneity. Another major reason for stratification is that, where marked differences exist between subgroups of the population (e.g., urban vs. rural areas), stratification allows flexible selection of the sample allocation and design separately for each subgroup. Stratification should be introduced only at the first stage of sampling. For this study, the sampling frame was stratified into 30 separate strata (districts). In each district the villages were ordered according to urban and rural classification to achieve the effect of implicit stratification which led to a proportional distribution of the sample villages in each stratum [12].

2.2.2 Systematic sampling:

Systematic sampling is the selection of sampling units at a fixed interval from a list starting from a randomly determined point (household for this case). Selection is systematic because selection of the first sampling unit determines the selection of the remaining households. The systematic sampling has the following advantages:

1. It is easier to perform.
2. It allows easy verification of the selection.
3. It provides a stratification effect with respect to the variables on which the frame is sorted and a proportional allocation.
4. Implicit stratification prevents unexpected concentration of sample points in certain areas.

The recommended household selection procedure is equal probability, systematic sampling. This procedure consists of selecting the sample households from the listing with a random start by the following criteria:

1. Calculate the sampling interval $I=N/n$, where N is the total number of HHs listed in the EA and n the number of HHs to be selected in the EA.
2. First selected sample HH is h if and only if: $(h-1)/L < Random \leq h/L$, Where Random is a random number between (0, 1).
3. Subsequent selected HHs are those having serial numbers: $h+(j-1)*I$, (rounded to integers) for j = 2, 3, …, n.

2.2.3 Sample Size and Allocation:

The fundamental goal of a survey is to come up with the same results that would have been obtained had every single member of a population been interviewed, and recall that the objective of a sample survey designs is to provide estimators with small variances at the lowest possible cost. After the sample size *n* is chosen, there are many ways to divide *n* into the individual stratum sample sizes, n_1, n_2 , …, n_L. Each division may result in a different variance for the sample mean. Hence, the main objective is to use an allocation that gives a specified amount of information at minimum cost. The best allocation scheme is affected by three factors:

1. The total number of elements in each stratum.
2. The variability of observations within each stratum.
3. The cost of obtaining an observation from each stratum.

Once the total sample size has been fixed, we need to appropriately allocate the sample to the various domains (areas) or, within domains, to the strata of interest. This allocation is aimed at strengthening the sample efficiency at the domain level [20]. At the first stage sampling, a sample size of 50 primary sampling units (PSUs) in each stratum in urban area and 40 PSUs in each stratum in rural area was drawn; which led to a selection of 1080 PSUs in rural and 150 PSUs in urban areas summing up to 1230 PSUs sampled from 30 strata countrywide. The selection within each stratum was done systematically using the probability proportional to size (PPS) from the ordered list of EAs in the sampling frame. At the second stage sampling, HHs selection were carried out using equal probability, systematic sampling by selecting 9 and 12 households in each sampled EA from each stratum in urban and rural areas respectively. That led to a selection of 1,350 HHs in urban and 12,960 HHs in rural areas summing up to 14,310 households all over the country.

2.2.4 Sample Selection Procedures:

At the first sampling stage the sample for EAs were selected within each stratum systematically with PPS from the list of EAs in the 2012 RPHC sampling frame. Within each stratum the following procedures were used:

1. Cumulate the measures of size (number of HHs) down the ordered list of EAs within the stratum. The final cumulated measure of size was the total number of HHs in the frame for the stratum (M_h).
2. To obtain sampling interval for stratum *h* (I_h), divide M_h by the total number of EAs to be selected in stratum h (n_h).
3. Select a random number (R_h) between 0.01 and I_h .

The sample EAs in stratum *h* was identified by the following selection numbers: $S_{hi} = R_h + [I_h * (i-1)]$ (rounded up) where i = 1,2,..., n_h. The i^{th} selected EA is the one with the first cumulated measure of size that is greater than or equal to S_{hi}. At the second sampling stage the systematic sample of m_{hi} households was selected from the HH listing for each sample EA using the following procedures:

1. All HHs in occupied housing units should be assigned a serial number from 1 to M'_{hi}, the total number of HHs listed.
2. To obtain the sampling interval for the selection of households within the sample EA (I_{hi}), divide M'_{hi} by m_{hi}, and maintain 2 decimal places.
3. Select a random number (R_{hi}) with 2 decimal places, between 0.01 and (I_{hi}).

The sample households within the sample EA was identified by the following selection numbers: $S_{hij} = R_{hi} + [I_{hi} * (j-1)]$, (rounded up), where j = 1, 2, 3,..., m_{hi} The j^{th} selected household is the one with a serial number equal to S_{hij} .

2.2.5 Design Weight:

A stratified two-stage sample design was used. Therefore, the sampling weights were calculated based on sampling probabilities that were computed separately for each sampling stage and for each stratum. The probabilities of selection for both sampled EAs and households were calculated using the following notations:

P_{1hi} : First stage's sampling probability of the i^{th} cluster (EA) in stratum h

n_h : Number of sample EAs selected in stratum h

M_{hi} : Expected number of HHs according to the 2012 Census frame for the i^{th} sample EA in stratum h

M_h : Expected number of HHs according to the 2012 Census frame for stratum h

The probability of selecting the i^{th} cluster in stratum h is calculated as follows: $P_{1hi} = \frac{n_h * M_{hi}}{M_h}$ **(1)**

P_{2hi} : Second-stage's sampling probability within the i^{th} cluster (households)

m_{hi} : Number of sample HHs selected in the i^{th} sample EA in stratum h .

M'_{hi} : Actual number of HHs from the new listing for the i^{th} sample EA in stratum h

The second stage's selection probability for each HH in the cluster is calculated as follows: $P_{2hi} = \frac{m_{hi}}{M'_{hi}}$ **(2)**

The overall selection probability (P_{hi}) of each HH in i^{th} cluster of stratum h is therefore the product of the selection probabilities: $p_{hi} = P_{1hi} * P_{2hi} = \frac{n_h * M_{hi}}{M_h} * \frac{m_{hi}}{M'_{hi}}$ **(3)**

The sampling weight for each household i th cluster of stratum h is the inverse of its selection probability:

$$W_{hi} = \frac{1}{P_{hi}} = \frac{M_h * M'_{hi}}{n_h * M_{hi} * m_{hi}} \quad \textbf{(4)}$$

2.3 Model:

Given the nature of variables in this study an ordered logistic regression (OLR) model was used to statistically measure the relationship between an ordinal dependent variable and a set of independent variables. An ordinal variable is a variable that is categorical and ordered, for instance, "poor", "good", and "excellent", which might indicate a person's current health status or HH's income status [2]. As a predictive analysis, OLR model describes data and explains the relationship between one dependent variable and two or more independent variables. One way to take account of the ordering is the use of cumulative probabilities, cumulative odds and cumulative logits. Considering k+1 ordered categories, these quantities are defined by:

$$P\ (Y \le i) = P_1 + \ldots + P_i \quad \textbf{(5)}$$

$$\text{odds}\ (Y \le i) = \frac{P(Y \le i)}{1\text{-}P(Y \le i)} = \frac{P_1 + \ldots + P_i}{P_{i+1} + \ldots + P_{k+1}} \quad \textbf{(6)}$$

$$\text{logit}\ (Y \le i) = \ln\left(\frac{P(Y \le i)}{1\text{-}P(Y \le i)}\right) \quad , \quad i=1,\ldots,k \quad \textbf{(7)}$$

The ordered logit model has the following characteristics:

- There is an observed ordinal variable, Y.
- Y, in turn, is a function of another variable, Y*, that is not measured.

- Y* is a continuous and unmeasured latent variable whose values determine what the Y equals.
- Y* has various threshold points (k) and the value on the observed variable Y depends on whether or not a particular threshold has been crossed.

In this study, the dependent variable is ordinal and independent variables are categorical (HH size, cow ownership, urban/rural and provinces). The dependent was portioned into three categories namely: Low, Middle and High income categories. Let consider for instance M = 3, then $Y_i = 1 \text{ if } Y^*_i \text{ is } \leq k_1$ **(8)**

$$Y_i = 2 \text{ if } k_1 \leq Y^*_i \leq k_2 \quad \textbf{(9)}$$

$$Y_i = 3 \text{ if } Y^*_i \geq k_2 \quad \textbf{(10)}$$

Once the score on unobserved latent variable Y* was k_1 or less, the score on Y would be 1; if the Y* score was between k_1 and k_2, Y would equal 2; and if the Y* score was above k_2, Y would equal 3. Briefly, the latent variable (Y*) can take on an infinite range of values which might then be collapsed into different categories of the observed ordinal variable (Y). The continuous latent variable Y* can be expressed as:

$$Y^*_i = \sum_{k=1}^{K} \beta_k X_{ki} + \varepsilon_i = Z_i + \varepsilon_i \quad \textbf{(11)}$$

Noting that, there is a random disturbance term which has a standard logistic distribution in this case. The OLR Model estimates part of the above: $Z_i = \sum_{k=1}^{K} \beta_k X_{ki} = E(Y^*i)$ **(12)**

Due to the random disturbance term, the unmeasured latent variable Y* can be either higher or lower than Z. The K βs and the M-1 ks are parameters that need to be estimated. Once they are estimated, using the corresponding sample estimates for each case then $Z_i = \sum_{k=1}^{K} \beta_k X_{ki}$

(13)

Note that there is no intercept term, then use the estimated M-1 cut-off terms to estimate the probability that Y takes on a particular value. The formulas are:

$$P(Y_i > j) = \frac{\exp(X_i\beta - k_j)}{1+[\exp(X_i\beta - k_j)]}, \; j=1, 2, \ldots, M\text{-}1' \quad \textbf{(14)}$$

This implies

$$P(Y_i = 1) = 1 - \frac{\exp(X_i\beta - k_1)}{1+[\exp(X_i\beta - k_1)]} \quad \textbf{(15)}$$

$$P(Y_i = j) = \frac{\exp(X_i\beta - k_{j-1})}{1+[\exp(X_i\beta - k_{j-1})]} - \frac{\exp(X_i\beta - k_j)}{1+[\exp(X_i\beta - k_j)]}, \; j=2, 3, \ldots, M\text{-}1 \quad \textbf{(16)}$$

$$P(Y_i = M) = \frac{\exp(X_i\beta - k_{M-1})}{1+[\exp(X_i\beta - k_{M-1})]} \quad \textbf{(17)}$$

In the case of M = 3, these equations are simplified to:

$$P(Y=1) = \frac{1}{1+[\exp(Z_i - k_1)]} \quad \textbf{(18)}$$

$$P(Y=2) = \frac{1}{1+[\exp(Z_i - k_2)]} - \frac{1}{1+[\exp(Z_i - k_1)]} \quad \textbf{(19)}$$

$$P(Y=3) = 1 - \frac{1}{1+[\exp(Z_i - k_2)]} \quad \textbf{(20)}$$

Hence, using the estimated value of **Z** and the assumed logistic distribution of the disturbance term, the ordered logit model can be used to estimate the probability that the unobserved variable **Y*** falls within the various threshold limits. The cumulative logistic model for ordinal response data is given by:

$$\text{logit}\,(Y \le i) = \alpha_i + \beta_1 X_1 + \ldots + \beta_{im} X_m \quad , \quad i = 1, \ldots, k \tag{21}$$

The OLR model has k model equations and one logistic coefficient β_{ij} for each category/covariate combination. Hence, the general cumulative logistic regression model contains a large number of parameters. However, in some cases a more parsimonious model is possible. If the logistic coefficients do not depend on i, we have only one common parameter c for each covariate. It follows that the cumulative odds are given by:

$$\text{odds}\,(Y \le i) = \exp(\alpha_i)\exp(\beta_1 X_1 + \ldots + \beta_{im} X_m) \quad , \quad i = 1, \ldots, k \tag{22}$$

which means that the k odds for each cut-off category i differ only with regard to the intercepts α_i; in other words, the odds are proportional [30].

2.3.1 Estimation of parameters:

In the aforementioned model, parameters were interpreted as follows: A positive coefficient indicates that a household having a particular trait (for example a household residing in the rural area) increases its chance or likelihood to be found in a particular category of the dependent variable (for example a higher income category), when other factors in the model are kept constant. A negative coefficient means that when other variables are held constant in the model, a household having a certain characteristic (let's say large household size) reduces its chance or likelihoods to fall in a particular category of the independent variable (middle income category).

2.3.2 Overall model fit:

A short introduction to what is meant by goodness-of-fit underlines the importance of assessing the adequacy of statistical models. The purpose of any model is to describe the relationship between a response and one or several covariates. Such models can be divided into a systematic component (the model function) and an error component (residuals). The error component consists of the deviations of the data from the systematic part. If these residuals are large then the model doesn't fit well and does not describe the data adequately. In that case, any conclusions drawn from this model are questionable. Hence, assessing goodness-of -fit plays a central role in the model building procedure and should be done before any hypotheses are tested [11]. The g***oodness-of-fit test***, in general, refers to measuring how well do the observed data correspond to the fitted model. It is used to compare the goodness of fit of two models, one of which (the null model) is a special case of the other (the alternative model) and it assesses the improvement of fit between the predicted and observed values on **Y** by adding the predictor(s) to the model. The test based on the likelihood ratio, which expresses how many times more likely the data are under one model than the other [3]. The goodness of fit of the model found in the output figure is interpreted using the following criteria: ***Significance level****:* after running a Fisher goodness of fit test, if the **P** value of the parameter was below 0.05 we concluded that the model is statistically significant and it proved that the independent variables have a significant impact on the independent variable.

2.3.3 F-test:

F-test is designed to test if two population variances are equal. It does this by comparing the ratio of two variances. All hypothesis testing is done under the assumption the null hypothesis ($\mathbf{H_0}$) is true. If $\mathbf{H_0}$ is true, then the F test-statistic can be simplified. If $\mathbf{H_0}$ is false, then we reject the $\mathbf{H_0}$ that the ratio was equal to 1 and the assumption that they were equal [16]. The F-test is performed using the steps below:

Step1: Grand Mean/ overall mean which is the total of all the data values divided by the total sample size. $\bar{X}_{GM} = \frac{\sum_i \bar{X}_i}{N}$, where N is the number of groups.

Step2: Between-group sum of squares is the variation due to interaction between the samples for Sum of Squares Between groups. The variation is comprised with the sum of the squares of the differences of each sample mean with the overall mean.

$SS(B) = n\sum(\bar{x}i - \bar{X}_{GM})^2$, where n is the number of data values per group.

Step 3: Between-group degree of freedom is one less than the number of groups $f_b = k\text{-}1$, where k is the number of group.

Step 4: Between-group mean square which is the between group sum of squares divided by its degrees of freedom. $MS_B = \frac{SS(B)}{f_b}$.

Step 5: With in-group sum of squares is the variation due to differences within individual samples denoted as $SS(W)$ for Sum of Squares Within groups. Each sample is considered independently, no interaction between samples is involved. Since each sample has degrees of freedom equal to one less than their sample sizes, and there are k samples, the total degree of freedom is k less than the total sample size: $df = N - k$. Hence, $SS(W) = \sum df * S^2$.

Step 6: Within group degrees of freedom is $f_w = a\ (n-1)$. Thus the within-group mean square value is: $MS_W = \frac{SS(W)}{f_w}$

Hence F-ratio is: $F = \frac{MS_B}{MS_W}$

2.4 Data cleaning:

Data cleaning deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data [4]. The data cleaning approach should satisfy several requirements. First of all, it should detect and remove all major errors and inconsistencies both in individual data sources and when integrating multiple sources. The approach should be supported by tools to limit manual inspection and programming effort and be extensible to easily cover additional sources. Furthermore, data cleaning should not be performed in isolation but together with schema-related data transformations based on comprehensive metadata [8]. The inconsistencies detected or removed might have been originally caused by human error (data entry clerks or enumerators), instrument error, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. For this study, the major data quality checks examined the consistency between questionnaire (particularly for questions related to agriculture section) and datasets; just to see whether the variables and value labels as well as routing instructions were the same both in questionnaire and datasets. The double checks were also carried out on skip patterns, outliers, valid values and missing values. Outliers in agricultural production and sales as well as income variables were examined so as to perform a suitable agricultural income analysis.

2.5 Treatment of outliers:

The outliers are defined as the most extreme or unusual observations which may include the observations maximum or observations minimum, or both, depending on whether they are extremely high or low. However, the observations maximum and minimum are not always outliers because they may not be unusually far from other observations [21]. Once the outlier is obviously due to incorrectly entered or measured data, first and foremost we don't ignore them as most of statistical methods are very sensitive to outliers or often they simply don't work [13]. Therefore, we have to deal with outliers' treatment. The three-sigma rule is a simple test for outliers if the population is assumed normal and as a normality test if the population is potentially not normal. In mathematical notation, three-sigma rule can be expressed as follows, where x is an observation from a normally distributed random variable, μ is the mean of the distribution, and σ is its standard deviation:

$$\Pr(\mu - \sigma \le x \le \mu + \sigma) \approx 0.6827$$
$$\Pr(\mu - 2\sigma \le x \le \mu + 2\sigma) \approx 0.9545$$
$$\Pr(\mu - 3\sigma \le x \le \mu + 3\sigma) \approx 0.9973$$

Using three-sigma rule as a test for outliers, analyst can compute the size of deviations in terms of standard deviations (σ), and compare this to expected frequency (μ); the points that fall more than 3 standard deviations from the norm are more likely to be outliers. If there are many points more than 3 standard deviations from the norm (outliers) then analyst simply replaces them by the mean of the values for that variable within the group containing the outliers [25]. In this study, extreme values in the agricultural production, sales and income variables were identified as observations more than

3 standard deviations larger than the mean of the variable in question or less than 3 standard deviations smaller than the mean of the variable. The outliers were expressed as a natural logarithm and calculated over crop and region.

2.6 Imputation:

Imputation is defined as the process of replacing missing data with substituted values. Because missing data can create problems for analyzing data, imputation is seen as a way to avoid pitfalls involved with list-wise deletion of cases that have missing values. Imputation preserves all cases by replacing missing data with an estimated value based on other available information [9]. The sample surveys whether conducted on people or on other types of units, almost always consist of a number of variables for which information is desired. In this array of data cells there are invariably some for which the data are missing [20]. Many ways can be used to deal with this problem; some researchers stated few of the basic techniques for replacing missing data:

1. Make a random choice from the values of the variable in question that is recorded in the sample.
2. Divide the sample into groups that may contain similar values of the variable in question and then select a value from the group that contains the missing value.
3. Replace the missing values by the mean of the values for that variable within the group containing missing value [19].

3. RESEARCH FINDINGS AND DISCUSSION

3.1 Distribution of the outcome variable:

The household agricultural income (HAI) variable was divided into 3 classes (Low, Meddle and High). According to the results presented in Table 3.1, the majority of household in Rwanda falls into medium range (53 %) of agricultural income whereas, 41.5% of household belongs to the lowest agricultural income. However, the highest agricultural income category constitutes only about 6% of households.

Table 3.1: Frequency of household agricultural income

Agricultural income categories					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Low Income	1034381	41.5	41.5	41.5
	Medium Income	1320313	53.0	53.0	94.5
	High Income	138351	5.5	5.5	100.0
	Total	2493044	100.0	100.0	

The distribution of HAI has significantly increased over the period of three years. This is proved by the findings in Table 3.2 whereby the mean household agricultural income raised up to 96,275 Rwf in 2013/14. Actually, compared to what was reported by NISR via the EICV3 agriculture thematic report carried out in 2010/11, there was a significant increase of 21,427 Rwf throughout this period (2010/11 and 2013/14).

Table 3.2: Mean of the HHs Agricultural income

```
Number of strata =      30        Number of obs    =    14419
Number of PSUs   =    1230        Population size  =  2493044
                                  Design df        =     1200
```

	Linearized			
	Mean	Std. Err.	[95% Conf.	Interval]
TOT_AGRINCOME	96274.66	1918.061	92511.53	100037.8

Referring to results presented in Table 3.3 the median of household agricultural income (HAI) is equivalent to 73,923 Rwf. Looking at the two tables (Table 3.2 and Table 3.3) we realize that the median of the data is less than the mean, this implies that there is a long right tail in the distribution of HAI or in other words the data are lopsided.

Table 3.3: Median of the HHs Agricultural income

Percentile estimation

TOT_AGRINC~E	Coef.	Linearized Std. Err.	z	P>|z|	[95% Conf.	Interval]
p50	73923.2	1150.678	64.24	0.000	71667.91	76178.48

The categorization of the HAI variable was set hypothetically by triangulation of data shown by the mean and median of agricultural income in the households. Whereby a household was qualified to be in the low agricultural income category if its annual agricultural earning is less or equal to 60,000 Rwf, in middle income category if its earning ranges between 60,001 and 250,000 Rwf and in high income category if it earns more than 250,000 Rwf.

3.2 Interpretation of the results in terms of coefficients of the ordered logistic regression model:

In interpreting the coefficients of the OLR model, the positive coefficients indicate that a household with a particular characteristic (residence area or household size) increase its likelihood to be found in a higher category of agricultural income while the negative coefficients indicates that a household having a particular characteristic (residence area or household size) reduces its likelihoods to fall in a lower category of agricultural income.

Table 3.4: Parameter Estimates of the ordered logistic regression model

```
Number of strata   =        30                  Number of obs      =     14419
Number of PSUs     =      1230                  Population size    = 2493044.3
                                                 Design df          =      1200
                                                 F(   7,   1194)    =    244.60
                                                 Prob > F           =    0.0000
```

AGRINCOME	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
HHSIZE						
HH with 5+People	-.6754154	.0387331	-17.44	0.000	-.7514075	-.5994232
HHCOW						
HH with a cow	1.544772	.0478744	32.27	0.000	1.450845	1.638699
REGION						
Rural	1.704588	.1238923	13.76	0.000	1.461518	1.947657
PROVINCE						
Southern	.6501165	.1805198	3.60	0.000	.295947	1.004286
Western	.3008212	.1802102	1.67	0.095	-.052741	.6543834
Northern	1.065212	.1859463	5.73	0.000	.7003957	1.430028
Eastern	.9765275	.183283	5.33	0.000	.6169366	1.336118
/cut1	1.850614	.1405361	13.17	0.000	1.57489	2.126338
/cut2	5.605366	.1538538	36.43	0.000	5.303514	5.907219

By looking at the top of Table 3.4, the results show that all 14,419 observations in the dataset are used in the analysis and with the application of weights the total number of households becomes 2,493,044 which reflect an estimation of all households in Rwanda in 2013/14. The Fisher test of 244.60 with a p value of 0.0000 indicates that the model as a whole is statistically significant, as compared to the null model with no predictors. The degrees of freedom of Fisher distribution used to test the t-statistic is defined by the number of predictors in the model which is 7 for this case. In the output, coefficients, their standard errors, t-tests and their associated p-values, and 95% confidence interval of the coefficients are shown. Almost all independent variables are statistically significant. Therefore, for HH size, as the p-value is equivalent to 0.000 which is less than 0.005, it is clear that for a one unit increase in household size, a decrease of -0.68 in the log odds for a HH to be in a higher category of agricultural income is expected, given that all other variables in the model are held constant. On the other hand, with a p-value of 0.000 we can conclude that for a one unit increase in cow ownership, an increase of 1.54 in the log odds for a HH to be in a higher category of agricultural income is expected. Similarly, as the

p-value is 0.000 which is less than 0.005 we can also conclude that for a HH to live in the rural area leads to an increase of 1.70 in the log odds for this household to be in a higher category of agricultural income. Referring to p-values obtained (almost equivalent to 0.000 which is less than 0.005) we can conclude that HHs staying in the Northern and Eastern provinces are more likely to earn the highest agricultural income (1.07 and 0.98 units increase in the log odds respectively) compared to other households living in the Southern province where , 0.65 units increase in the log odds. However, as the p-value is 0.095 in the Western province which is greater than 0.05 we can conclude that the HHs staying in this province have less chance of being in the highest agricultural income category (0.30 units increase in the log odds). The model used in this study is written as:

$$AGRINCOME_j = \beta_1 HHSIZE_j + \beta_2 HHCOW_j + \beta_3 REGION_j + \beta_4 Southern + \beta_5 Western + \beta_6 Northern + \beta_7 Eastern + Error_j \quad \textbf{(23)}$$

$$AGRINCOME_j = -0.68 HHSIZE_j + 1.54 HHCOW_j + 1.70 REGION_j + 0.65 Southern + 0.30 Western + 1.07 Northern + 0.98 Eastern + Error_j \quad \textbf{(24)}$$

3.3 Hypothesis testing for Ordered Logistic Regression Model:

In this study, the hypotheses are formulated in the following manner:

The null hypothesis is: $H_0: \beta_1 = 0$ and $\beta_2 = 0$ and ... and $\beta_7 = 0$ So, under the null hypothesis, there is in fact no effect of the predictor variables in the model (i.e.: all of the regression coefficients in the model are equal to zero).

The alternative hypothesis is: $H_1 : \beta_1 \neq 0$ or $\beta_2 \neq 0$ or... or $\beta_7 \neq 0$ Under the alternative hypothesis, there is in fact effect of the predictor variables in the model (i.e.: at least one of the regression coefficients in the model is not equal to zero). To analyze the goodness of fit of the model a Fisher test was run and a P value of 0.00 was obtained; as this value is less than 0.05 we reject the null hypothesis ($\mathbf{H_0}$) and accept the alternative hypothesis ($\mathbf{H_{1)}}$ which implies that this model is statistically significant, meaning that the observed data perfectly correspond to the fitted model. In other words the selected predictors have a statistically significant association with the outcome variable.

3.4 Interpretation of the results in terms of predicted probabilities in OLR model:

As we are dealing with an OLR model, the predicted probabilities can be also obtained and used to interpret the results, which are usually easier to understand than the coefficients. The findings in Table 3.5 explain how the probabilities of HHs' association to each category of agricultural income vary as the cow ownership variable changes in the HH when other variables are held constant in the model. The HHs without any cow are more likely to be found in the lowest category of agricultural income than HHs with at least one cow (0.53 as opposed to 0.19), about as likely to report they are in the middle category of agricultural income (0.45 against 0.72) and about as likely to say they are in the highest category of agricultural income (0.02 versus 0.09). Briefly, the probability of being in the lowest category of agricultural income becomes very high for the HHs not having a cow; while the probability of being either in the middle or in the highest category of agricultural income increases as the HHs owns at least one cow.

Table 3.5: Distribution of households agricultural income, according to cow ownership

Agri-income Classification	Delta-method					
	Margin	Std. Err.	z	P>z	[95% Conf. Interval]	
Low Income						
HH without cow	0.528	0.008	64.11	0.000	0.512	0.544
HH with cow	0.193	0.007	26.1	0.000	0.178	0.207
Middle Income						
HH without cow	0.451	0.008	57.8	0.000	0.436	0.466
HH with cow	0.718	0.006	111.62	0.000	0.705	0.731
High Income						
HH without cow	0.020	0.001	18.77	0.000	0.018	0.023
HH with cow	0.089	0.004	21.81	0.000	0.081	0.097

With reference to the HH size variable, the results displayed in Table 3.6 indicate how the likelihoods of households' linkage to every group of agricultural income behave as the household size variable changes in the case where other factors in the model are kept constant. The results show that when other variables are held constant in the model, households with size of five persons or more are more likely to fall in the lowest category of agricultural income than households with size ranging between one and four persons (0.49 as opposed to 0.33), about as likely to report they are in the middle category of agricultural income (0.49 against 0.63) and less than half as likely to say they are in the highest category of agricultural income (0.02 versus 0.05). In other words this means that, the probability of being in the lowest category of agricultural income becomes very high for the households with a large HH size. The data also expose that the probability of being either in the middle or in the highest category of agricultural income increases as the HH size decreases.

Table 3.6: Distribution of households agricultural income, according to household size

Agri-income Classification	Delta-method					
	Margin	Std. Err.	z	P>z	[95% Conf. Interval]	
Low Income						
HH Size of 1-4 People	0.330	0.008	40.37	0.000	0.314	0.346
HH Size of 5+ People	0.491	0.009	54.21	0.000	0.474	0.509
Middle Income						
HH Size of 1-4 People	0.625	0.008	82.27	0.000	0.610	0.640
HH Size of 5+ People	0.485	0.009	56.67	0.000	0.468	0.502
High Income						
HH Size of 1-4 People	0.045	0.002	21.05	0.000	0.041	0.050
HH Size of 5+ People	0.024	0.001	18.92	0.000	0.021	0.026

Referring to the area where a household lives, the results presented in Table 3.7 reveal how strongly the probabilities of households'attachment to each category of agricultural income change. The HHs living in urban area are more than twice as likely as HHs living in rural area to be found in the lowest category of agricultural income (0.74 as opposed to 0.34), about as likely to say they are in the middle category of agricultural income (0.26 against 0.62) and one quarter as likely to report they are in the highest category of agricultural income (0.01 versus 0.04). In other words, the probability of being in the lowest category of agricultural income becomes very high for the HHs residing in urban area. On the contrary, the probability of being either in the middle or in the highest category of agricultural income increases for the HHs living in rural area.

Table 3.7: Distribution of households agricultural income, according to urban/rural

Agri-income Classification	Delta-method					
	Margin	Std. Err.	z	P>z	[95% Conf. Interval]	
Low Income						
HH in Urban Areas	0.736	0.021	34.57	0.000	0.694	0.778
HH in Rural Areas	0.336	0.008	41.52	0.000	0.320	0.352
Middle Income						
HH in Urban Areas	0.256	0.020	12.53	0.000	0.216	0.296
HH in Rural Areas	0.620	0.008	82.51	0.000	0.605	0.634
High Income						
HH in Urban Areas	0.008	0.001	8.57	0.000	0.006	0.010
HH in Rural Areas	0.044	0.002	20.64	0.000	0.040	0.048

When it comes to provincial level, the findings presented in Figure 3.1 indicate how the probabilities of households' membership to each category of agricultural income change according to whether the households live in a given province. The HH in Western Province present the high probability of being in low category of agricultural income (0.49) and least probability of being either in the middle or in the highest category of agricultural income (0.49 and 0.02) compared to other Provinces. Similarly, households in Southern Province recode the high probability of being in low agricultural income category (0.40), and low probabilities of being either in the middle or in the high category of agricultural income (0.56 and 0.03). The data also reveal that, households in the Northern Province rank first with the highest probabilities of being either in the middle or in the highest category of agricultural income (0.64 and 0.05) compared to all other Provinces.

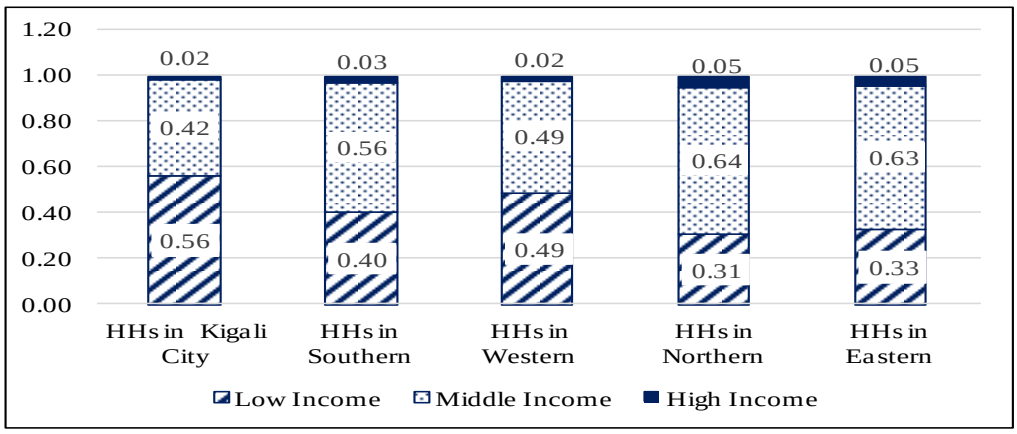

Figure 3.1: Presentation of households' agricultural income, according to residential province

4. CONCLUSION AND RECOMMENDATION

4.1 Conclusion:

The ordered logistic regression model was used to statistically measure the relationship between an ordinal dependent variable and a set of independent variables. After carrying out statistical analysis we come up with the following conclusions:

The findings confirmed that there is a substantial relationship between the household agricultural income and cow ownership as the HHs owning at least one cow have higher probability of being either in the middle or in the highest category of agricultural income while those not having any cow have higher chances of being in the lowest category.

Again the data showed a significant linkage between household's agricultural income and its size, as the bigger the number of household members the higher chances for that household to be found in the lower agricultural income, on the other hand, the HHs with small size are more likely to be either in the middle or in the highest category of agricultural income.

The results also highlighted that there is a considerable connection between the household's agricultural income and its residential area since the likelihood for a HH to be either in the middle or in the highest categories of agricultural income rises when that household resides in a rural area. The situation reverses when the HH lives in urban area as its probability of being in the lowest agricultural income bracket gets very high.

Analysis demonstrated that there is a relationship between household agricultural income and its residential province. Compared to all other Provinces, households in the Northern Province have the highest chance of being found either in the middle or in the highest agricultural income category. However, residents of the Western Province are more likely to earn the least agricultural income.

Finally, agricultural income tended to be either over-estimated or under-estimated in analysis, which resulted into many outliers in the data. For that reason, the choice of respondent should be considered (most knowledgeable household member).

4.2 Recommendations:

Based on the summary of findings from this study, the following recommendations have been made as possible ways to open door for further research and show the government of Rwanda about the priority area of intervention:

1. We highly recommend the government of Rwanda with its partners to strengthen and continue enhancing the provision of cows to poor and vulnerable families as the results strongly supported that, cow ownership would offer a pathway out of poverty and a vital means to economic growth for a large number of households keeping livestock.
2. We strongly recommended that, the government of Rwanda and other institutions in charge of family planning to frequently continue taking up the policy of family planning as the findings reiterated that, family size influences a lot the household's quality of life particularly the household's agricultural income.

3. We encourage the researchers to continue doing further analysis to find out the reasons why the Western province continues to lag behind in terms of well being particularly for farming households.
4. We also suggest that the government of Rwanda through its policies of production intensification and farmers' income boosting across the country to put more efforts and consider that Province as priority area of intervention.
5. We recommend that NISR should use multiple respondents in interviewing households since the use of a single respondent might result in a loss of accuracy in capturing individual income by household members that the survey respondent should not observe.

REFERENCES

[1] Agresti, A. (2002) Categorical Data Analysis, Second Edition. Hoboken, New Jersey: John Wiley & Sons, Inc.

[2] Anderson, J. A. 1984. Regression and ordered categorical variables (with discussion). Journal of the Royal Statistical Society, Series B 46: 1–30.

[3] Brown CC. On a goodness-of-fit test for the logistic model based on score statistics. *Comm Stat* A 1982; 11: 1087-105.

[4] Chaudhuri, S., Dayal, U.: An Overview of Data Warehousing and OLAP Technology. ACM SIGMOD Record 26(1), 1997.

[5] Christiaensen, L.J., Demery, L., 2007. Down to Earth: Agriculture and Poverty in Africa. World Bank, Washington, DC.

[6] Dethier, J.-J., Effenberger, A., Agriculture and development: A brief review of the literature. Econ. Syst. (2012), doi:10.1016/j.ecosys.2011.09.003.

[7] EICV3 agriculture thematic report 2010/2011.

[8] Galhardas, H.; Florescu, D.; Shasha, D.; Simon, E.: Declaratively cleaning your data using AJAX. In Journees Bases de Donnees, Oct. 2000. http://caravel.inria.fr/~galharda/BDA.ps.

[9] Gelman, Andrew, and Jennifer Hill. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, 2006. Ch.25

[10] Grafarend, Linear and Nonlinear Models: Fixed Effects, Random Effects, and Mixed Models, Walter de Gruyter, 2006, p. 553.

[11] Harrell FE Jr, Lee KL, Mark DB. Multivariate prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; 15:361-87.

[12] ICF International. 2012. *Demographic and Health Survey Sampling and Household Listing Manual*. MEASURE DHS, Calverton, Maryland, U.S.A.: ICF International.

[13] John Tukey. His seminal work Tukey (1977) "Exploratory Data Analysis".

[14] Johnston, B.F., Mellor, J.W., 1961. The role of agriculture in economic development. The American Economic Review 51, 566– 593.

[15] Liao, T. F. (1994) Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models. Thousand Oaks, CA: Sage Publications, Inc.

[16] Lomax, Richard G. (2007). *Statistical Concepts: A Second Course*. p. 10. ISBN 0-8058-5850-4.

[17] Maxwell, S. and Fernando, A. (1989) 'Cash crops in developing countries: the issues, the facts, the policies', World Development, 17(11): 1677–1708.

[18] Mellor, J.W., 2001. Faster more equitable growth – agriculture, employment multipliers and poverty reduction. Agricultural Policy Development Project Research Report 4, Cambridge, MA.

[19] Paul S. Levy, Stanley Lemeshow Sampling of Populations-Methods and Applications 3rd Ed. (Wiley Series in Survey Methodology) 1999.

[20] Richard L. Scheaffer, III William Mendenhall, R. Lyman Ott, Kenneth G. Gerow Elementary Survey Sampling 2011.

[21] Maddala, G.S. (1992)."Outliers "Introduction to Econometrics (2nd ed.) New York: MacMillan. pp. 88–96 [p. 89]. ISBN 0-02- 374545-2.

[22] Integrated Household Living Conditions Survey 2013/2014 Main indicators report| National Institute of Statistics of Rwanda.

[23] Poverty Profile Report 2013/2014| National Institute of Statistics of Rwanda.

[24] Statistical Yearbook, 2015 Edition, November 2015| National Institute of Statistics of Rwanda.

[25] EICV3 Thematic report_ Income| National Institute of Statistics of Rwanda.

[26] *Schaum's Outline of Business Statistics. McGraw Hill Professional. 2003. p. 359,* and in *Grafarend, Erik W. (2006). Linear and Nonlinear Models: Fixed Effects, Random Effects, and Mixed Models. Walter de Gruyter. p. 553.*

[27] Schultz T.W. (1945). "Agriculture in an unstable economy". New York: McGraw Hill, 1945.

[28] Schultz, T.W. ,1964. Transforming Traditional Agriculture. Yale University Press, New Haven, CT.

[29] William G. Cochran1977. Sampling Techniques Third Edition, New York: Wiley.

[30] Peter McCullagh. Journal of the Royal Statistical Society. Series B (Methodological), Volume 42, Issue 2. (1980), 109-142.