International Journal of Advanced Statistics and Probability, 4 (1) (2016) 31-35

International Journal of Advanced Statistics and Probability

Website: www.sciencepubco.com/index.php/IJASP
doi: 10.14419/ijasp.v4i1.5657
Research paper

Binary logistic regression to estimate household income efficiency. (south Darfur rural areas-Sudan)

Sofian A. A. Saad [1]*, Amin I. Adam [2], Afra H. Abdelateef [3]

[1] *University of Nyala, Faculty of Economics, Department of Economics, Nyala-Sudan*
[2] *Omdurman Islamic University, Faculty of Economics, Department of Statistics, Omdurman-Sudan*
[3] *Sudan University of Science & Technology, Faculty of Science, Department of Statistics*
**Corresponding author E-mail: sofianabuelbacher@yahoo.com*

Abstract

The main objective behind this study is to find out the main factors that affects the efficiency of household income in Darfur rejoin. The statistical technique of the binary logistic regression has been used to test if there is a significant effect of fife binary explanatory variables against the response variable (income efficiency); sample of size 136 household head is gathered from the relevant population. The outcomes of the study showed that; there is a significant effect of the level of household expenditure on the efficiency of income, beside the size of household also has significant effect on the response variable, the remaining explanatory variables showed no significant effects, those are (household head education level, size of household head own agricultural and numbers of students at school).

Keywords: *Income efficiency; Household Head (HH); Factors affecting income; Odds Ratio (OR).*

1. Introduction

Not necessarily that all people in certain region, area, society and or a country to be have an equal level of income, since there is innate differences in the capabilities of person to another, in addition to the variation in the distribution of natural resources from one area to another further to many other factors that help some people to get a larger size of income while depriving others. This paper aims to identify the most important factors that affect the adequacy of household income in the Darfur region (largest region in Sudan), which has seen an era of armed conflict lasted for more than ten years leads to a significant impact on the sources of income of individuals and families in the Region.

Household income has been touted as one of the most effective measures of families' well-being. 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.). 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.

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 helps related development actors to fairly distribute the general income and government subsidies among the whole society.

During the first decade of this century all Darfur states (Sudan) especially South Darfur state has faced horrible militant crisis by some military movement targeting the civilians in the region which leads to serious social and economic collapse especially in terms of the livelihood for the affected people, these situation together with the scarcity of the natural resources at that time forces a majority of the people for losses their main sources of income and then hindering their ability to have a better type of life. However, in this paper we would like to see the main factors that affect their income sufficiency depending on household face to face data gathering.

Research problem

The problem of the study concentrate mainly on how can we come out by the main effective variables that affected the efficiency of 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) To study and discuss the importance of the application of binary 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.

Types of variables

- Dependent variable: Level of income
- Independent variables are;
  - i) Household size (mean number of family members)
  - ii) Household head (HH) level of education
  - iii) Level of Expenditure
  - iv) Size of own Agricultural land use
  - v) Student at school

Table 1: Variables Formulation

Variables	Codes	
Income level(Dependent)	≤ 1300 SDG Insufficient (0)	>1300 SDG Sufficient (1)
Household size	≤ 5 (0)	>5 (1)
Agric land size	≤ 5 feddan (0)	> 5 feddan (1)
Level of Expenditure	≤ 50 pound (0)	>50 pound (1)
HH education level	Illiteracy (0)	Educated (1)
Students at school	≤ 5 (0)	> 5 (1)

Hypotheses of the study:

The study will adopt the following hypotheses:-

1) The significance of regression coefficient B_1 for the expenditure level is equal to zero (H_0: $B_1 = 0$.).
2) The significance of regression coefficient B2 for the agricultural land size is equal to zero (H_0: $B_2 = 0$).
3) The same are for the rest of the other variables coefficients (i.e., student at school, household head education level and household size).

Sample framework

The sample framework of the study shall be the south Darfur rural Societies with significant insufficient sources of income.

Sampling and data gathering

It is obvious that fresh and reliable answers to the critical policy questions can only be obtained by interviewing a sizeable and representative sample of households carefully drawn from an up-to-date sampling frame of the population study (South Darfur State).

This paper will examine the following questions:

1) What are the main variables that will determine the efficiency of income?
2) Do all explanatory variables have the same effects on the response variable? And also we would like to see,
3) If there a significance relationship between the response variable and the predictors.

To examine the study questions, a sample of size 136 household head have been drawn from the origin population using sampling clustering techniques with full probability selection.

2. Literature review

Household income is often equivalised to account for differences in living needs between families of different size and composition. Very rare studies have been done regarding the household income since there are huge factors affecting the determinants of the household income.

[1] In their study for the income in rural areas; they reflect that the determination of the income mainly in the rural areas will be influences by many factors that affecting production in these areas, such as labour education level, environmental condition, the size of the labour force and land development.

In case of Sudan where the current study takes place in a biggest region in the country, many other factors will also be have enormous effects on the determination of rural household income, mainly; conflicts, war, poor economic policy and fragile economic structures.

[2] "Determinants of Income Distribution in Nigerian Economy". The study aimed to explore the most important factors that serve to identify the degree of inequality in income distribution in the economy of Nigeria for the period (1977-2005), Gini coefficient was calculated, which measures the degree of inequality in the distribution of income in a country. The finding shows that Gini equal to 0.52, which is considered very high index coefficient, indicates the magnitude of the disparity in income among households in the Nigerian society. The study used the applied method and standard analysis according to cointegration approach to analyze the relation between these factors and income distribution. The main variables that used in the study are the rate of unemployment, inflation, gross domestic product and manufacturing expenses. The study has shown that there is a very significant impact of these variables on increment of Gini coefficient. The study also proved that the growth rate and government spending on health sector has a negative relationship with the Gini coefficient reduces the severity of the disparity in income between individuals, while the high unemployment rate and increase spending level on education sector and the high rate of inflation had a positive relationship with the Gini coefficient. The study demonstrated a long-term relationship between the Gini coefficient and the independent variables, and recommended that the government should formulate appropriate policies to protect the operating system and increase the level of spending on health and education sectors to guarantee equity and fairness of income distribution between household and individuals.

[3] In a study "Does educational achievement help to explain income inequality"? He examines income inequality among four education classes: those with less than a High School Diploma, those with a High School Diploma, those with some college, and those with a college degree or more. The Gini Coefficients is computed for all four groups for the years 1950 to 2009 using the Decennial Census and the 2009 American Community Survey. The result indicated that income inequality was initially driven by those with less than a high school education, was passed onto those with a high school diploma, and in recent years has greatly increased due to those with a college education. In addition to that, wage inequality was greater among the college educated than among the other groups.

In his study (Factors affecting the income inequality) to identify the most important determinants that lead to disparities in the distribution of income in Jordan,[4] had studied some factors such as demographic, economic, social, health and cultural factors to see if they have an impact on the disparity in the distribution of income among families and individuals. The findings was adopted using the regression analysis as well as the correlation analysis between the different variables with the greatest impact on the phenomenon under the study, he found that there is a linear correlation between the independent variables and the dependent variable. In addition to that, the correlation analyses between the independent variables and the dependent variable showed that the disparities in income distribution which measured by Gini coefficient might affected by many demographic, social and economic factors and there is a significant positive relationship between the disparity in income distribution and the following estimated variables; the average of household size, percentage of urbanization, the dependency ratio, and average of individual income. The study also showed that there was no significant relationship between the disparity in income distribution and health, cultural factors.

Other study conducted by [5] in Indian rural areas, using some statistical techniques (i.e. ordinary least square, maximum likelihood) they found that there is a clear evidence that there a relationship between a person being poor or rich and the area where he or she lived, beside that they also found that there is a big relation between a household with low income and poverty and the agricultural performance. However, factors that lead to high level of agricultural production must be given more consideration in order to alleviate poverty and increase the level of rural income.

Some factors also found have direct effects on income inequalities, these are; the macroeconomic factors, which are inflation, unemployment, the size of government's expenditure, external debt and foreign reserves, changes in the exchange rate, and other factors. In other research study ("as stated by Chris Crowe [6]") the research offers an explanation for the positive cross-sectional relationship between income inequality and inflation. In addition to that other research shows that unemployment has inequality increasing effects, because high unemployment worsens the situation of those at the bottom of income distribution. The influence of the government's expenditure depends on its composition, mainly on the share of social transfers in public expenditure

Regarding the relationship between income inequality and human capital, ("as found by Gary Becker [8]") people with high level of education are more likely to gain more money than those with low level of education, because education increases their ability and skills to be more competitive in the labour market as well as increases the level of their productivity.

In their study to find the relationship between income distribution and level of education as one of the most important human capital, [7] present empirical evidence on how education is related to the income distribution. Their findings indicate that higher education attainment and more equal distribution of education play a significant role in making income distribution more equal.

It is clear that from the selective previous studies regarding the factors affecting rural household income, that there is various factors lead to inefficiency of income in low income rural areas. In

order to raise the level of income for household in these areas countries should adopt comprehensive economic policies together with taking into consideration the particular surrounding circumstances for each area.

3. Methodology

Sometimes researchers find themselves in a position to carry out an analysis of qualitative variables for models in order to find out the relations between these variables. The regular way to do the analysis when the variables were qualitative is to put them in a form of dummy variables consist of two or more values depending on the nature of the study. However, the ordinary least squares (ols) in this case will not provide accurate results and the postulated findings give fake shape for the estimated parameters. Therefore, they cannot be relied upon to predict the relations between these variables, that is because putting the dependent variable in that way increase the possibility of falling in the problem of heterogeneity (variance random errors in the estimated models were not equal)

In order to get logical and accurate results for functions having qualitative variables, it is better to use one of the following techniques:

1) Discriminant function analysis, in a condition that all the independent variables in the model must be continuous and normally distributed.
2) Logit Regression Model, which depend on the Cumulative Logistic Probability Function. And
3) Logistic regression analysis

To get the findings of this study and since some of the data were in a categorical type and the dependent variable is qualitative, hence the method of logistic regression will be used.

The logistic regression is one of the most useful statistical techniques that can be used to estimate the probability of a categorical outcomes variable.

Logistic regression analysis can be of two types:

1) Binary logistic regression. And
2) Multinomial logistic regressions.

The binary logistic regression used, when the dependent variable is dichotomous with two possible outcomes, for example (yes or no) (success or failure), while the multinomial logistic regression is applied when the dependent variable is dichotomous with more than two possible outcomes and the rest of the independent variables should be continuous , categorical (nominal and/or order) and scale. Not like the normal ordinary least squire the logistic regression needs no assumptions should be required to estimate the model.

After estimating the model it is crucial to calculate the probabilities of the model predictors in regard with the dependent variable outcomes, to see the possible chances for each values (i.e. 0,1 or yes/no) of the dependent variable to what extent it would be tend to happened. Odds and odds ratio will also be calculated which gives informative information and explain the relations between the dependent variable and predictors.

$Odds = e^{a+bx}$ Where (e) is the exponential term. And the,

$$OddsRatio(OR) = \frac{\text{Odds for certain response outcome}}{\text{Odds for the other response outcome}}$$

The logistic regression model of binary response always takes the form as:

$$\pi(x) = \frac{e^{(\alpha+\beta x)}}{1+e^{(\alpha+\beta x)}} = \frac{1}{1+e^{-(\alpha+\beta x)}} \quad (1)$$

Which gives the probability of the dependent variable equaling to one of the response variable out that have been selected by the researcher (0 vs 1 or yes vs no or dead vs alive).

Equivalently, the log odds, called the logit , has the linear relationship

$$\log it\left[\pi(x)\right] = \log\frac{\pi(x)}{1-\pi(x)} = \alpha + \beta x \quad (2)$$

This equation gives us the estimated model for the variables under the study, then Interpreting α, Odds, Probabilities, and linear approximations for the whole model.

4. Results and conclusions

4.1. Model validation

Table 2: Classification

Observed		Predicted incsats insufficient	Percentage Correct sufficient	
incsats	insufficient	57	16	78.08
	sufficient	15	48	76.19
Overall Percentage			77.21	

Where incsats in table (2) refers to income satisfaction.

The above classification table (2) from output result summarizes the observed group and the predicted group classification. It is obvious that, the overall correctly specified group percentage is 77.21% which is good result that tells us the model well fitted the data. To consolidate these results also we getting the receiver operating characteristic curve (ROC) which is also used to indicate the sensitivity and specificity for all possible cutoff points. The idea behind the ROC curve is to calculate the area under it. The ROC cover an area range from 0.5 to 1.00, the closer the value under the curve to 1.00 it is an evidence of better fit. From the data 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) which mean that the logistic regression classifies the group significantly better than by chance.

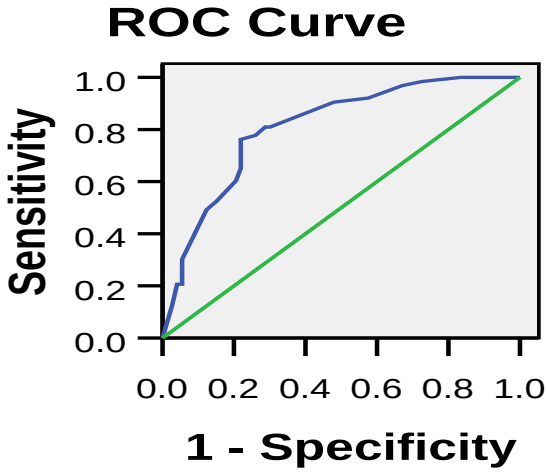

Fig. 1: Receiver operating characteristic Curve

Table 3: Hosmer & Lemeshow Test

Step	Chi-square	df	Sig.
1	5.216	7	0.634

The Hosmer-Lemeshow tests the null hypothesis that predictions made by the model fit perfectly with observed group memberships, so from the table No (3) since the value of chi-square is 5.216 hence there is no reason to reject the null hypothesis.

4.2. Goodness of fit

Regarding the overall goodness of fit of the model, we used the log likelihood ratio which follows chi-square distribution. As we see that from the table No (3) the chi-square is equal to (44.449) with (p-value = 0.000) indicate that the overall model is significant.

Table 4: Omnibus tests of model coefficients

	Chi-square	df	Sig.
Model	44.449	5	0.000

4.3. Model inference

Table 5: Variables in the equation

	B	S.E	Wald	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
						Lower	Upper
expn	2.287	.423	29.200	.000	9.847	4.296	22.572
landsize	-.014	.459	.001	.976	.986	.401	2.427
schstu	.637	.536	1.412	.235	1.891	.661	5.408
edulevl	-.049	.446	.012	.913	.953	.398	2.281
hhsize	.278	.176	4.252	.035	1.321	.445	3.918
Constant	-1.612	.548	8.663	.003	.199		

Variable(s) entered on step 1: expn, landsize, schstu, edulevl, hhsize

From the information in table (5) we can fit the logistic regression model as:

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = -1.612 + 2.287x1 - .014x2 + .637x3 - .049x4 + .278x5$$

Where X1 is expenditure, X2 is Agric land size, X3 is student at school, X4 is household head education level and X5 is the household size.

We can now use this model to predict the odds for the independent variables those appear to be have a significant relation with the dependent variable, that a subject of a given predictor will likely to be has sufficient income or insufficient. The odds prediction equation is; $Odds = e^{a+bx}$

For the first repressor (expenditure) If our subject is a person whose level of expenditure is ≤ 50 pound (expn = 0), the $Odds = e^{-1.612+2.287(0)} = 0.199$ That is, a person with an expenditure level ≤ 50 pound is only 0.199 as likely to have sufficient income level as he is to have an insufficient income. If our subject a person whose level of expenditure is > 50 pound (expn =1), the $Odds = e^{-1.612+2.287(1)} = 1.964$; that is a person with expenditure level > 50pound is 1.964 times more likely to have sufficient income. The same will be done for the second significant independent variable (i.e., hhsize).

If we would like to see the probabilities for each group concerning the significant regresor this will be done through the conversion of the odds as follow;

For a person with expenditure level ≤ 50 pounds, $\pi = \frac{odds}{1+odds} = \frac{0.199}{1+0.199} = 0.17$ That is, the model predicts that 17% of persons with expenditure level ≤ 50 pounds will have sufficient income. For a person with level of expenditure is > 50 pounds $\pi = \frac{odds}{1+odds} = \frac{1.964}{1+1.964} = 0.66$ That is, the model predicts that 66% of person with expenditure level ≤ 50 pounds men will have sufficient income.

The Variables in the Equation output also gives us the Exp (B). This is better known as the odds ratio predicted by the model. This odds ratio can be computed as follows:

For our model and for the first variable (expn) $Odds = e^{2.287} = 9.845$ which mean that the model predicts that the odds of being has sufficient income are 9.845 times higher for a person with expenditure level > 50 pounds than they are for a person with expenditure level ≤ 50 pounds. For the person with expenditure level > 50 pounds, the odds are 1.964, and for the person with expenditure level ≤ 50

Pounds they are 0.199. The odds ratio is $\frac{1.964}{0.199} \approx 9.845$

For the model coefficients inference, it is clear that from the table (No 4) three variables showed no significance effects on the dependent variable, those are landsize, schstu and edulevl with significance level equal to 0.976, 0.235 and 0.913 respectively, where as the other two variables expenditure (expn) and household size (hhsize) showed significance effect on the determination of income sufficiency (dependent variable) with Wald statistic (= 29.200) and (sig level = 0.000) for the predictor variable expenditure and Wald statistic (= 4.252) and (sig level = 0.035) for the predictor variable household size.

Usually we are not interested in the significance level of the intercept (the consistent variable).

5. Conclusion

The findings reflect the importance of the application of binary logistic regression in the analysis, since the nature of the elected data was categorical.

The logistic regression analysis showed that not all the variables have the same effects on the dependent variable.

Two of the variables were significantly affected the dependent variable (i.e., level of expenditure & the size of household).

Three of them reflect no significant effects on the dependent variable (i.e., numbers of students at school, household head education level and agricultural land size).

Acknowledgement

Special thanks to my supervisors Dr. Amin Ibrahim, Dr. afra Hashim and to my colleague Dr. Elsiddig Alsadig for their greatest efforts that help me more in the publication of this paper.

References

[1] Paul A. Samuelson, William D. Nordhaus. (1992). Economics. 14th edition. ISBN 0071128115 (pbk.) : 007054879X

[2] A.A. Awe , Olawumi Ojo Rufus. (2005), Determinants of Income Distribution in the Nigeria Economy: 1977-2005. International Business and Management, Vol. 5, No. 1, 2012, pp. 126-137

[3] Checchi, D. (2000). Does educational achievement help to explain income inequality? Working paper, JOURNAL of Sociology Mind, Vol.1 No.4, October 14, 2011.

[4] Kharabsheh. Factors affecting Income Distribution in Jordan, published M.Sc thesis, Derasat journal, Jordan 2011

[5] Richard Palmer-Jones and Kunal Sen. (2006). Agricultural Economics. Volume 34, Issue 3, pp 207–345. Article first published online: 21 APR 2006 | DOI: 10.1111/j.1574-0864.2006.00121.x. The Journal of the International Association of Agricultural economists.

[6] Christopher Crowe. 'Inflation, Inequality and Social Conflict', CEP Discussion Paper No. 657 http://cep.lse.ac.uk/pubs/download/dp0657.pdf

[7] Gregorio, Jose and Lee, Jong-Wha, 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

[8] Gary S. Becker, 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

[9] Chia (2008, p.233) finds that family income constraints do matter in determining whether children participate regularly in sporting activities. Drewnowski and Specter (2004)

[10] Lantz PM, House JS, Lepkowski JM, Williams DR, Mero RP, Chen J. Socioeconomic factors, health behaviors, and mortality: Results from a nationally representative prospective study of US adults. JAMA. 1998; 279(21):1703-1708. http://dx.doi.org/10.1001/jama.279.21.1703.
[11] Agresti, A (2002), "Categorical Data Analysis", John Wiley & Sons, Inc. http://dx.doi.org/10.1002/0471249688.
[12] Agresti, A (2007), "An Introduction to Categorical Data Analysis", John Wiley & Sons, Inc. http://dx.doi.org/10.1002/0470114754.
[13] Chatterjee S, and Hadi A (2006) "Regression Analysis by Example", John Wiley & Sons. http://dx.doi.org/10.1002/0470055464.
[14] http://en.wikipedia.org/wiki/Logit
[15] Atia (2004, 372-375) "The linear probability model",
[16] www.asu.edu.jo/asu/userfiles/file/HumanitiesSeries-pdf/...15-1.../3.pdf.