The International Journal Of Engineering And Science (IJES)
|| Volume || 3 || Issue || 6 || Pages || 01-05 || 2014 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Logistic Regression: A Paradigm for Dichotomous Response Data

[1]Ogunfiditimi, Franklin (Ph.D) and [2]Oguntade, Emmanuel

[1] Dept of Mathematics, University of Abuja, FCT, Nigeria

[2] Dept of Statistics, University of Abuja, FCT, Nigeria

--ABSTRACT---

Poverty is seen as a disease that has infected many homes in different countries of the world which need urgent and immediate treatment for Millennium Development Goals to be achieved or actualised. This article presents the theoretical basis for binary response data set, as well as the empirical results of the analysisof demographic data obtained from various households in Gbagyi Community of the Federal Capital Territory, Abuja, Nigeria.The study examines the prevalence rate and possible causes of poverty in Dobi, Gwako and Bako communities in Gwagwalada Area Council with the application of Logistic Regression Model. The empirical analysis reveal that socioeconomic status and level of education of household head are inversely related. It also shows a strong association between poverty and level of income of household head. While age, size, assets of household head and other demographic variables considered show various levels of insignificance in the estimated model. Ageing increases the likelihood of poverty while literacy, family size, sex decreases the chance .Based on the empirical results, the researcher recommends the establishment of more public schools so as to increase their level of education, provision of basic social amenities as well as encourage family planning among rural dwellers so as to serve as booster of socio economic status and prevent poverty trap in time.

KEYWORDS: Logit Model, Odds Ratio, Probability Model, Dichotomous data, Poverty, poverty trap.

Date of Submission: 03 June 2014 Date of Publication: 05 July 2014

I. INTRODUCTION

Logistic regression is part of a category of statistical models called generalized linear models which includes ordinary regression, Analysis of Variance (ANOVA), Analysis of Covariance (ANCOVA) and lognormal regression (Agresti, 1996).Logistic regression is a standard tool for modelling effects and interactions with binary response data (Park & Hastie, 2007). It makes possible the prediction of a discrete outcome from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these using the most parsimonious model. Logistic regression or linear probability model as popularly called gives the conditional probability that an event will occur given the values of the regressors as well as providing the knowledge of the relationships and strengths among variables, (Wright, 1995; Hyun & Ditto, 2006; Park & Hastie, 2007).Healy (2006) opined that Logistic regression is based on the log of the odds of a particular event occurring with a given set of observations. Its underlying principles are based on probabilities and the nature of the log curve. The only assumptions of Logistic regression are that the resulting logit transformations are linear, the dependent variable is dichotomous and that the resultant logarithm curve doesn't include outliers. Hence Normality assumptions such as homogeneity of variance, observations and disturbance terms are normally distributed and all normality tests are invalid and Ordinary Least Square (OLS) assumptions break down due to dichotomous nature of dependent data.Logistic Regression is preferred by many researchers in the analytical fields due to its robust practical nature , intuitive assumptions and its ability to produce a predictive representation of the real world situations,(Healy,2006).

Logit model has been used extensively in analysing growth phenomena, such as population, GDP and money supply (Krammer, 1991); It has also featured in manufacturing and health related studies, (Healy, 2006); Recreational activities (Hyun &Ditton, 2006);Examination results(Saha.2011);determinants of Poverty (Achia et al,2010)etc The researcher apply a common theme of the theory of Logit models by placing an individual into distinct categories or groups rather than along a continuum As a result of the various usefulness of this all encompass model that has a non linear relationship between the response and the predictor variables. It got applications in various field of studies which includes epidemiology studies, demography, social sciences among others.

Hence the research work aims at using axiom and the concept in question to find the possible statistical relationship between poverty and annual household income, level of education of household head, physical assets, sex of household head, and size of household among other variables possibly contributing to poverty in Gbagyi community of Federal Capital Territory Abuja Nigeria.

II. THEORETICAL AND CONCEPTUAL FRAME WORK

Poverty is a global phenomenon and is dynamical in nature with many facets, (World Bank, 2000; World Development Report, 2001; NEEDS, 2004; Achia et al, 2010).

Poverty is the inability to retain a minimal standard of living measured in terms of basic consumption needs or some income required for satisfying them (World Bank, 2000).

According to contemporary English dictionary poverty is defined as state of being poor, lack quality or state of being inferior.

Poverty is a multidimensional phenomenon (World Bank, 2000). Its various dimensions includes : lack of opportunities , lack of empowerment, lack of security, low access to health care facilities and other social infrastructural facilities which make the poor citizens vulnerable to diseases, violence and so on.

Poverty is dynamic and has many dimensions (NEEDS,2004).People may get into poverty as a result of natural disasters or health problems, lack of access to credit , or the lack of natural resources .Poor people are more likely to live in rural areas, be less educated and have larger families than the rest of the population.

According to NEEDS (2004), Poverty has many causes, all of which reinforce one another. One source of poverty is lack of basic amenities /services such as clean water, education and health care. Another is lack of assets, such as land, tools, credits and supportive networks of friends and family. A third is lack of income, including food, shelter, clothing and empowerment.

Households are poor and inequality in the distribution of incomes, assets, amenities and other social services prevails, (NEEDS, 2004; Okunmadewa et al., 2005).

See Achia et al, 2010; Saha, 2011 for extensive literature and referencing on Poverty, demographic variables, determinants of poverty in developing countries and logistic regression models among others.

There are so many assertions made about poverty in the continent of Africa where Scholars believed that Poverty killed faster than HIV/AIDS and some identified poverty as the possible cause(s) of unlimited uprising, unrest and civil disobedience in the various regions.

Here, we seek to determine the trends and key determinants of poverty using various demographic variables in some hamlets of Gbagyi Community in the Federal Capital Territory Abuja, Nigeria, West Africa with the application of Logistic Regression Model.

2.1 THE MODEL

Regression model for binary response variables is used to describe the population proportions of occurrence of an event. The population proportion of successes represents the probability $p(y=1)$ for a randomly selected subject. This probability varies according to the values of the explanatory variables.Models for binary data which are dichotomous in nature assume a binomial distribution which are well described in Hollander & Wolfe (1973:Pp15); Whittle, (1976); Krzanowski, (1998:Pp18); Gujarati, (2004:Pp583); Spiegel et al, (2004); and Awogbemi&Oguntade, (2012) for the dependent variable. Let y represent a dichotomous random variable, and then the binary response variable y has two categories denoted by 1 and 0.That is, the dependence variable can take the value 1 with a probability of success λ, or the value 0 with probability of failure $1-\lambda$.This type of variable is called a Bernoulli or binary variable

Let $p(y=1)=1-p(y=0)=\lambda$ (1)

Where λ is defined by the equation (2)

$$\lambda = \frac{\ell^{(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n)}}{1+\ell^{(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n)}} \qquad (2)$$

That is (2) can be succinctly written as

$$\lambda = \frac{\ell^{(\alpha+\sum_{i=1}^{n}\beta_i x_i)}}{1+\ell^{(\alpha+\sum_{i=1}^{n}\beta_i x_i)}} \qquad (3)$$

Where α serves as the bench mark for the equation and β_i is the coefficient of the exogenous variables x_i for $i=1,2,...,n.$

The first task in the model estimation is to transform the independent variable and determine the coefficients of the independent variables, (Healy, 2006).The basic logistic regression analysis begins with logit transformation of the dependent variable through utilization of maximum likelihood estimation. This is done using what is popularly known as Odds Ratio. The odds ratio for an event is represented as the probability of the event outcome divided by one minus probability of event outcome.
The odds ratio is given by

$$Odds = \frac{\lambda(x)}{1-\lambda(x)} = \ell^{(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n)} \qquad (4)$$

Since from (3)

$\lambda = \frac{\upsilon}{1+\upsilon}$,where $\upsilon = \ell^{(\alpha+\sum \beta_i x_i)}$

therefore $Odds = \frac{\lambda(x)}{1-\lambda(x)} = \frac{\upsilon}{1+\upsilon}.\frac{1}{1-\frac{\upsilon}{1+\upsilon}}$

But, $1-\frac{\upsilon}{1+\upsilon} = \frac{1+\upsilon-\upsilon}{1+\upsilon} = \frac{1}{1+\upsilon}$

Therefore, $Odds = \frac{\upsilon}{1+\upsilon}.\frac{1+\upsilon}{1} = \upsilon$

Where : $\lambda(x)$ is the probability of success i.e. an event occurring,and
$1-\lambda(x)$ is the probability of failure i.e. an event not occurring.

$\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$ represent the regression model.

Hence equation (4) can be transformed into an alternative form of logistic regression equation by taking the Naperian logarithm of the odds ratio popularly known as logistic transformation (Logit) to obtain equation (5)

$$\log it\left[\lambda(x)\right] = \ln\left[\frac{\lambda(x)}{1-\lambda(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \qquad (5)$$

(For elaborate details and extensive referencing on Logistic regression models, See: Krammer, (1991), Krzanowski, (1998:Pp189); Gujarati (2004), Healy, (2006), Park & Hastie, (2007)to mention but a few scholars).

WALD TEST OF SIGNIFICANCE FOR THE MODEL PARAMETER

To determine the significance of the independent variables we can use either the Wald statistic or the likelihood ratio test, (Healy, 2006).The Wald statistic is a method to test whether the coefficients are significantly different from zero. It is used to test the statistical significance of each coefficient (β) in the model. A Wald test calculates a Z statistic:

$$z = \frac{\hat{\beta}_i}{se_{\beta_i}}$$

This Z value is then squared, yielding a Wald statistic with a chi-square distribution.

2.3 LIKELIHOOD RATIO TEST OF INDEPENDENCE

The likelihood-ratio test uses the ratio of the maximized value of the likelihood function for the full model (L_1) over the maximized value of the likelihood function for the simpler model (L_0). The likelihood-ratio test statistic equals:

$$z = \frac{\hat{\beta}_i}{se_{\beta_i}} - 2\log(\frac{L_0}{L_1}) = -2\left[\log\left(L_0\right) - \log\left(L_1\right)\right] = -2\left(L_0 - L_1\right)$$

This log transformation of the likelihood functions yields a chi-squared statistic.

III. EMPERICAL ANALYSIS

3.1 THE DATA

Random samples of 250 indigenes in Gbagyi communities of Gwagwalada Area Council (GAC) Abuja, Nigeria, were selected for the study. The selected hamlets includes: Gwako, Dobi, and Passo -Gbagyi communities in FCT Abuja. The data gathered were collected using self designed and administered questionnaires and the process was consistence with Hyun and Ditto, 2006.

3.2 RESULTS AND DISCUSSION

Table 1: Model information

Table 1: Model information

								95.0% C.I.for EXP(B)	
		B	S.E.	Wald	Df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Age	.011	.013	.656	1	.418	1.012	.963	1.016
	Size	-.038	.046	.705	1	.401	.962	.880	1.052
	Sex(1)	-.333	.453	.538	1	.463	.717	.295	1.744
	Education(1)	-.171	.004	7.405	1	.040	.843	.833	.853
	Landlord(1)	.203	.315	.412	1	.521	1.224	.660	2.272
	Income(1)	1.258	.335	14.096	1	.000	3.517	1.824	6.780
	Migrant(1)	-.402	.312	1.660	1	.198	.669	.363	1.233
	Marriage(1)	.475	.321	2.182	1	.140	1.608	.856	3.018
	Constant	.879	.882	.992	1	.319	2.408		

a. Variable(s) entered on step 1: Age, Size, Sex, Education, Landlord, Income, Migrant, Marriage.

From table 1 above, variables income and education significantly determine the poverty level in the said communities while all other variables including the drift are not statistically significant at 5% level of significance. Thus, the regression model reduced to equation (6).

$$Y/1 = 1.258X_{income(1)} - 0.171X_{education(1)} \qquad (6)$$

Hence, the fitted Logistic regression Model is

$$\ln Y/1 = \frac{\ell^{1.258X_{income(1)} - 0.171X_{education(1)}}}{1 + \ell^{1.258X_{income(1)} - 0.171X_{education(1)}}}$$

Log Likelihood is obtained as:
-2LogLikelihood= Reduced Model-Full Model
=288.925-269.209=21.482

The result shows that in Gbagyi communities under study, the level of education of household head is inversely related with the incidence of poverty in the community. Thus, an increase in education attainment has an important impact on reducing the probability that a household is poor as the variable is statistically significance as shown in Table1.Likewise, income as a demographic variable is also significant in determining poverty status in the estimated logistic regression model; hence, level of income as a measure of per capita income of household head of the said populace is highly associated with social economic status of Gbagyi Communities. While poverty and other demographical factors considered in the said Gbagyi communities are not a function of each other as variables have nothing to do with prevalence of poverty in the community going by the outcome of the analysis of the data gathered.The odd ratio of literacy to illiteracy is 0.843; odd ratio of high income to low income is 3.517. The computed odds ratio of data obtained in Table 1 indicates that ageing increases the probability of poverty, family size decrease probability of poverty, Literacy decrease probability of poverty, low income increases probability of poverty. A sample population has a decrease probability of having poverty if the estimated odds ratio of its variable(s) is less than one and if the odds ratio is greater than one then there will be an increased probability of poverty prevalence.

Table2:Hosmer and Lemeshow Goodness of Fit Test

Step	Chi-square	Df	Sig.
1	15.136	8	.057

From Hosmer and Lemeshow test, the data provides a good fit to the model estimates, as theHosmer statistics gives a non significance Chi- square value of 15.136 at 5% level of significance

Table 3: Omnibus Tests of Model Fit Coefficients (Likelihood Ratio test)

		Chi-square	Df	Sig.
Step 1	Step	21.482	8	.006
	Block	21.482	8	.006
	Model	21.482	8	.006

-2log Likelihood = 21.482 which is significance at 0.05 level of significance.This implies that at least one of the parameters is not equal to Zero and the model is well fitted. Thus, the full model is significant as shown by -2LogLikelihood statistic.

REFERENCES

[1] Achia T.N.O., Wangombe, A, and Khadioli N. (2010): ’*A logistic Regression Model to Identify Key Determinants of Poverty Using Demographic and Health Survey Data’* in European journal of Social Sciences. Volume 13:Number1, pp38-45.

[2] Agresti, A. (1996): *An Introduction to Categorical Data Analysis*. New York: John Wiley and sons, Inc.

[3] Awogemi C .A. and Oguntade E.S. (2012): *Element of Statistical Methods*. USA: LAMBERT Academic Publishing.

[4] Chung H, Flaherty B.P. and Schafer J (2006): ’*Latent Class Logistic Regression: Application to Marijuana Use and Attitude among High School Seniors”* in Journal of Royal Statistical Society.Vol.169, part4, pp.723-743.

[5] Fox, J. (2010): *Notes: Logit and Probit Models*. NewYork: SPIDA.

[6] Gujarati, D.N. (2004): *Basic Econometrics*, 4th Ed. New York: Tata Graw – Hill Publishing Co. Ltd.

[7] Healy, L.M. (2006): *Logistic Regression: An Overview.* Easter Michigan College of Technology

[8] Hollander, M. and Wolfe, D.A. (1973): *Nonparametric Statistical methods.* London: John Wiley and Sons.

[9] Krzanowski W.J. (1998): *An Introduction to Statistical Modelling.* London: Arnold Publishers.

[10] Hyun, W. Y.andDitton R. B. (2006): ‘*Using Multinomial Logistic Regression Analysis to Understand Anglers willingness to Substitute other Fishing Locations*’*in*Proceedings of the 2006 North eastern Recreational symposium.GTR-NRS-P-14:248-255.

[11] Krammer J.S. (1991): *The Logit Model for Economists*. London: Edward Arnold Publishers.

[12] Park M. Y and Hastie T. (2007): *Penalised Logistic Regression for Detection Gene Interactions*. Department of Statistics, Stanford University.

[13] NEEDS (2004): *NEEDS NIGERIA: Meeting every one’s Needs*.National Planning Commission. Abuja-Nigeria

[14] Saha G. (2011): ‘Applying Logistic Regression Model to the Examination Results Data’in Journal of Reliability and Statistical Studies. Vol. 4 Issue2:105-117.

[15] Spiegel, M.R., Schiller J., and Srinivasan, R. A.(2004) *Probability and Statistics, second edition*. New York: McGraw-Hill Publishing Company.

[16] Whittle P. (1976): *Probability*. London: John Willy &Sons.

[17] World Bank (2000): Development Report 2000/01, attacking Poverty. Washington DC.

[18] Wright, R. E.(1995):Logistic Regression.Pages217-224 in Hyun W. Y.andDitton R. B.(2006): ‘*Using Multinomial Logistic Regression Analysis to Understand Anglers willingness to Substitute other Fishing Locations*’in Proceedings of the 2006 Northeastern Recreational symposium.GTR-NRS-P-14:248-255.

[19] www.nigerian.stat.gov.ng.