A Logistic Regression Model to Identify Key Determinants of Poverty Using Demographic and Health Survey Data

Thomas N O Achia[1]
School of Mathematics, University of Nairobi
E-mail: achia@uonbi.ac.ke

Anne Wangombe
School of Mathematics, University of Nairobi
E-mail: wachiraanne@hotmail.com

Nancy Khadioli
School of Mathematics, University of Nairobi
E-mail: kkanancy@yahoo.com

Abstract

This study examines the determinants of poverty in Kenya. While most of the studies done on poverty determinants rely on the income, expenditure and consumption data, The data used in this study comes from the Demographic and Health Surveys, (DHS). The principal component analysis was used to create an asset index which gave the social economic status of each household. A Logistic regression was estimated based on this data with the SES (that is poor and non-poor) as the dependent variable and a set of demographic variables as the explanatory variables. The results presented in this paper suggest that the DHS data can be used to determine the correlates of poverty.

Keywords: Principal components analysis, Logistic regression

Introduction

The measurement and analysis of poverty have traditionally relied on reported income or consumption and expenditure as the preferred indicators of poverty and living standards. Income is generally the measure of choice in developed countries while the preferred metric in developing countries is an aggregate of a household's consumption expenditures, Sahn and Stifel (2003). The choice of expenditures over income is influenced by the difficulties involved in the measuring income in the developing countries. Similarly with the expenditure data the limitation is the extensive data collection which is time- consuming and costly as stated by Vyas and Kumaranayake (2006).

In this paper, we construct an asset index using Principal Component Analysis (PCA) from asset ownership variables in the Kenya Demographic and Health Survey (2003) and use logistic regression to identify key determinants of poverty in Kenya.

The use of demographic and health survey data to the measure of poverty is not unique. Filmer and Pritchett (2001) used Demographic and Healthy Survey data to show that the relationship between wealth and enrollment in school can be estimated without income or expenditure data, by using household asset variables. PCA provided acceptable and reliable weights for an index of asset to serve as a measure for wealth. In the four countries examined; India, Indonesia, Nepal and Pakistan this

[1] Partially supported by EAUMP_ISP (Sweden)

approach produced reasonable results. Filmer and Pritchett (1999) and Filmer (2002) explored how education attainment profile differed by wealth and gender in more then 35 countries using the DHS data. Sahn and Stifel (2000) employed demographic and healthy survey data in an analysis of poverty in nine African countries, they used principal component analysis to construct asset index. Booysen (2002) used demographic and healthy survey to measure differences in socioeconomic status of South Africa households. The asset index used represented a comparable indicator of poverty in South Africa.

Most of the studies done on poverty in Kenya relies on the expenditure and consumption data and thus use the poverty line computed from the Kenya Intergrated Household Budget Survey data using the of cost of basic needs method. While literature on poverty measurement is by now relatively developed and abundant, there are very few studies dealing with finding the determinant or causes of poverty. In their study, Mwabu et al. (2000) used regression analysis and identified the following variables as the key determinants of poverty: size of household, places of residence(urban or rural), level of schooling and livestock.

The most recent study on the determinant of poverty was done by Oyugi et al (2000). In their study they used Probit Model to analysis the Welfare Monitoring survey (1994) data. The predictors (household characteristics) used in the study included holding area, livestock unit, the proportion of household members able to read and write, household size, sector of economic activity (agriculture, manufacturing/industrial sector or wholesale/retail trade), source of water for household use, and off-farm employment. The result showed that almost all the variables used were important determinants of poverty.

Rodriguez and Smith (1994) used a logistic regression model to estimate the effect of different economic and demographic variables on the probability of a household being in poverty in Coasta Rica. The source of the data was from National Household- Income (1986). Their results showed that poverty was higher for the household whose heads had lower level of education.

An asset-index approach to the measuring of poverty is one alternative to income or consumption and expenditure. This approach although lacking data on income, consumption and expenditure, collects information on ownership of a range of durable assets which include; car/track, refrigerator, television, radio, bicycle, telephone and solar power, housing characteristic which includes material of dwelling floor, roof and toilet facilities and access to basic services which includes electricity supply, source of drinking water.

2. Methodology

2.1. The Data

The data used to analyze the poverty is taken from the 2003 Demographic and Health Surveys (DHS) for Kenya. The survey covered both rural and urban populations. The survey collected information relating to demographic and detailed information on asset ownership, access to public services and housing characteristics. A household was defined as a person or a group of people related or unrelated to each other, who live together in the same dwelling unit and share a common source of food.

The Demographic and Health Surveys utilized a two-stage sample design. The first stage involved selecting sample points (clusters) from a national master sample maintained by Central Bureau of Statistics (CBS) the fourth National Sample survey and Evaluation Programme (NASSEP) IV. In 2003, a total of 400 clusters, 129 urban and 271 rural, were selected. From these clusters, the desired sample of households was selected using systematic sampling methods.

2.2. Computation of a Poverty Index using Principal components analysis

We applied PCA to create an asset index based on data from the KDHS (2003). The KDHS (2003) included information regarding the ownership of durable goods, housing characteristic, access to services along with basic demographic information concerning household size and composition. Using

PCA, we first recoded the household variables into dichotomous variables, distinguishing between household that own the particular asset or for which a particular statement about access to services is true and one that do not own the asset or for which the statement is not true. Hence all variables take on a value of zero or one. The only variable that is included in the PCA as a continuous variable is the number of household members sharing a room for sleeping purposes.

The PCA is a multivariate statistical technique used to reduce the number of variables without losing too much information in the process. The PCA technique achieves this by creating a fewer number of variables which explain most of the variation in the original variables. The new variables which are created are linear combinations of the original variables. The first new variables will account for as much as possible of the variation in the original data.

Given *p* variables $X_1,...,X_p$ measured in *n* households, the *p* principal components $Z_1,...,Z_p$ are uncorrelated linear combinations of the original variable, $X_1,...,X_p$, given as

$$Z_1 = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p$$
$$Z_2 = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p$$
$$\vdots$$
$$Z_p = a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p$$

This system of equations can be expressed as $\boldsymbol{z}=\boldsymbol{A}\boldsymbol{x}$, where $\mathbf{z}=(Z_1,...,Z_p)$, $\boldsymbol{x}=(X_1,...,X_p)$ and $\boldsymbol{A}$ is the matrix of coefficients.

The coefficient of the first principal component, $a_{11},...,a_{1p}$, are chosen in such a way that the variance of Z_1 is maximised subject to the constraint that $a_{11}^2 + ... + a_{1p}^2 = 1$. The variance of this component is equal to λ_1, the largest eigenvalue of $\boldsymbol{A}$. The second principal component is completely uncorrelated with the first component and has variance equal to λ_2, the largest eigenvalue of $\boldsymbol{A}$. This component explains additional but less variation in the original variable than the first component subject to the same constraint. Further, principal components (up to the maximum of *p*) are defined in a similar way.

Each principal component is uncorrelated with all the others and the squares of its coefficients sum to one. The principal component analysis involves finding the eigenvalues and eigenvectors of the correlation matrix.

2.3. Logistic Regression Model

To identify key determinants of poverty we first computed a dichotomous variable indicating whether the household is poor or not. That is,

$$SES = \begin{cases} 1, & \text{if household is poor} \\ 0, & \text{otherwise} \end{cases}$$

where SES denotes social economic status.

On the basis of Pearson's Chi-square statistic, we determine whether the predictors age of household, size of household, educational level of the household head, type of residence(rural or urban), ethnicity and religion were associated with the poverty index, SES.

We then used a Logistic regression model, given by

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + + \ldots + \beta_6 X_6,$$

where $X_1,\ldots,X_6$ were the predictor variables age of household, size of household, educational level of the household head, type of residence(rural or urban), ethnicity and religion, respectively and p denoted the probability that the household was poor, was used.

The forward selection, backward elimination and stepwise (logistic) regression methods were determine automatically which variables to add or drop from the model. The conditional options use a computationally faster version of the likelihood ratio test.

3. Results

3.1. The Poverty Index

Table 1 shows all variables used in the construction of the asset index and the result of the PCA.

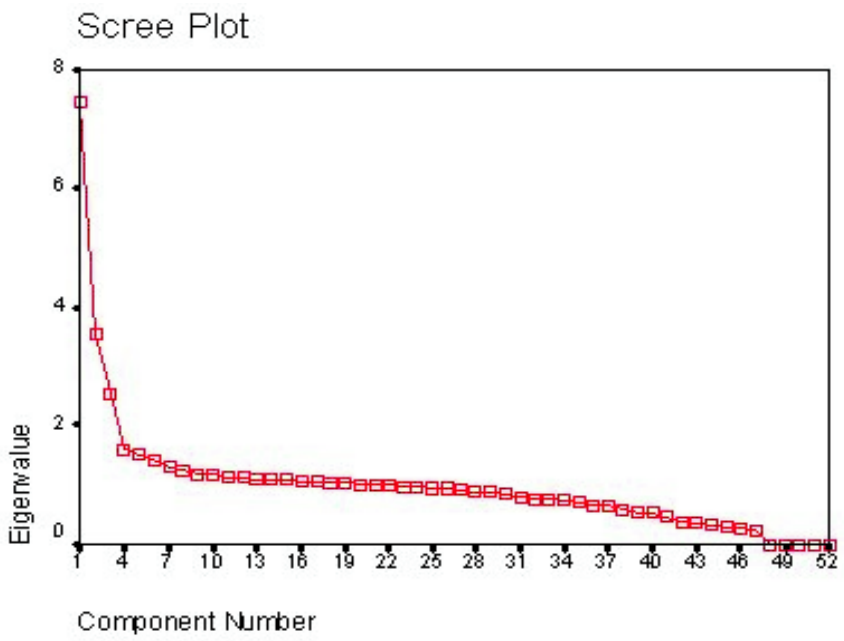
Table 1: Principal component score

Variable	Mean	SD	Component score	*N*	Variable	Mean	SD	Component score	*N*
Source of drinking water					**Type of roof material**				
Piped into dwelling	0.13	0.33	0.091	7906	Grass	0.22	0.42	-0.058	7906
Piped into compound	0.13	0.33	0.030	7906	Tin	0.00	0.05	-0.003	7906
Tap water	0.11	0.32	-0.002	7906	Iron sheet	0.65	0.48	0.026	7906
Well into compound	0.02	0.13	-0.003	7906	Concrete	0.05	0.21	0.052	7906
Public well	0.06	0.24	-0.021	7906	Tiles	0.04	0.20	0.073	7906
Covered well in the compound	0.05	0.22	0.000	7906	Others	0.02	0.13	-0.010	7906
Covered public well	0.06	0.23	-0.013	7906	**Cooking fuel**				
Spring	0.12	0.33	-0.028	7906	Electricity	0.01	0.07	0.023	7906
River	0.20	0.40	-0.041	7906	Gas	0.07	0.26	0.087	7906
Pond	0.01	0.11	-0.010	7906	Bogas	0.00	0.05	0.013	7906
Dam	0.04	0.21	-0.024	7906	Kerosene	0.13	0.34	0.043	7906
Rain	0.02	0.14	-0.001	7906	Coal	0.00	0.02	0.002	7906
Bottle	0.00	0.06	0.020	7906	Charcoal	0.16	0.37	0.022	7906
Others	0.05	0.23	-0.002	7906	Firewood	0.63	0.48	-0.098	7906
Sanitation facility					**Other durable goods**				
Flush toilet	0.16	0.37	0.110	7906	Has electricity	0.22	0.42	0.113	7906
Latrine	0.59	0.49	-0.054	7906	Has radio	0.77	0.42	0.042	7906
Ventilated	0.08	0.27	0.013	7906	Has television	0.28	0.45	0.100	7906
No facility\ bush	0.16	0.37	-0.046	7906	Has refrigerator	0.09	0.29	0.097	7906
Others	0.01	0.07	0.001	7906	Has bicycle	0.29	0.46	-0.009	7906
Type of floor material					Has motorcycle/scooter	0.01	0.09	0.010	7906
Earth or sand	0.56	0.50	-0.100	7906	Has car/truck	0.09	0.28	0.077	7906
Planks	0.01	0.01	0.023	7906	Has telephone	0.20	0.40	0.105	7906
Palm	0.00	0.02	-0.001	7906	Solar power	0.04	0.19	0.013	7906
Polished	0.01	0.10	0.037	7906	No. of rooms for sleeping	2.18	1.24	0.024	7906
Asphalt	0.01	0.09	0.020	7906					
Ceramic	0.02	0.13	0.041	7906					
Cement	0.37	0.48	0.066	7906					
Carpet	0.02	0.13	0.030	7906					
Other	0.00	0.07	0.017	7906					

The results of PCA indicate that the first principal component explains 14.3% of the variation in the original variables and each subsequent component explains a decreasing proportion of variance.

The screeplot in Figure 1 shows the proportion of variance explained by each principal component and indicates that the first four components would sufficiently explain the original variables.

Figure 1: Scree Plot



In the construction of the social economic index, only the factor score (that's the eigenvectors) of the first principal component are used.

3.2. Cross Tabulations

This section presents social economic status cross-tabulated by characteristics of the household; Education, household size, religion, region, age of household head, ethnicity and household own land. The asset index derived from the DHS data was employed to calculate estimate of the headcount poverty index for Kenya. The asset index at the *40*-th percentile is employed as the poverty line.

Table 2: Values of Pearson's χ^2- statistic on cross-classifying demographic characteristics with SES

Explanatory variable	χ^2- value	df	p-value
Type of place of residence	1767.69	1	<0.001
Highest education level	1397.38	3	<0.001
Religion	252.094	4	<0.001
Ethnicity	1146.8649	14	<0.001
Number of household members	140.7	2	<0.001
Age of household head	33.25	3	<0.001
Region	1505.11	7	<0.001

The results indicate that there is association between SES and the following predictor variables: Ethnicity; Religion; Number of household members; highest education level; Age of household head and Type of place of residence.

The distribution of households by ethnicity and social economic status results show that there is a difference in the very poor according to ethnicity with the Kikuyu having the least number of the poor represented by 18.3% while the Turkana having the highest number represented by 87.6%

According to the region (Provinces), the result show that Nairobi has the lowest number of the very poor, while North Eastern has the highest case of 86.2%. The distribution of households by religion and the SES result show that the other type of religion have lower cases of poverty which account for 20.7% while those with No religion taking the highest number of the very poor which

account for 74.2%. According to level of education, the result shows that the groups with the highest poverty cases have no education while those with the higher education have lower cases of poverty. The result correlates with the KIHBS (2005/2006) report. The result showed that in Kenya, the level of education of household head is inversely related with incidences of poverty both in rural and urban areas. We classified the household size using three household size categories (1-3, 4-6 and 7+). The result shows that poverty is highest for household with 7 or more members and lower for households of smaller sizes. The result correlates with the KIHBS (2005/2006) report which also used the same household size categories. According to the residence, the result shows that most cases of the very poor are in rural areas. The result correlates with the KIHBS (2005/2006) report and with the household head, The result shows that the cases of poverty increases with age of household head. The results compare favorably with the KIHBS (2005/2006). The poverty estimates are not directly comparable, given that different poverty lines, equivalence scale, time and data set are employed in estimating the headcount poverty index.

3.3. Logistic Regression Analysis

The final model that was fit to the data was given by

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7$$

where X_1 is Education, X_2 is the Residence, X_3 is Ethnicity, X_4 is the Region, X_5 is the Religion and X_6 is the age of the household. This was arrived at using a forward stepwise selection method.

The results indicate that there are higher levels of poverty in communities further from the national capital, Nairobi. Worst cases of household poverty relative to Nairobi are found to North Eastern province (OR: 17.46, 95%CI: 11.64-26.18), a semi-arid region, lacking in road and any proper infrastructure, followed by Nyanza province, a region marked with high HIV prevalence (25%) and poor infrastructure.

Religion was found to significantly explain household socio-economic status, adjusting for various socio-cultural and demographic factors. Households whose heads a Protestants (or other Christians) were more likely to be poor compare with those headed by Catholics (OR: 1.52, 95% CI: 1.31-1.76), households headed by a Muslims (OR: 2.47, 95% CI: 1.70-3.60) were more likely to be poor compared those headed by Catholics.

The results also indicated ethnicity and type of residence significantly explained the distribution of poverty. Rural households were more likely to poor as compared to Urban households (OR: 24.00, 95% CI: 19.81-29.09). It was also apparent that nomadic communities, like the Somali, Turkana and Maasai, had higher prevalence of poverty than other communities in the country.

Table 3: Table of Effects for the best fitting Logistic regression model

Variable	B	SE(B)	Exp(B)	p-value	Variable	Par Est	SE(B)	Exp(B)	p-value
Education					**Religion**				
No education	Reference				Catholic	Reference			
					Protestants/ other				
Primary	-0.63	0.1	0.53	< 0.001	Christians	0.42	0.07	1.52	< 0.001
Secondary	-1.31	0.12	0.27	< 0.001	Muslims	0.91	0.19	2.47	< 0.001
Higher	-1.32	0.16	0.27	< 0.001	No	2.87	0.26	17.71	< 0.001
					Others	**-0.26**	**0.58**	**0.77**	**0.648737**
Age					**Type of residence**				
15-19	Reference				Urban	Reference			
20-24	-0.01	0.09	0.99	0.944	Rural	3.18	0.1	24	< 0.001
25-29	0.01	0.1	1.01	0.881	Ethnicity				
30-34	-0.12	0.11	0.89	0.259	Embu	Reference			
35-39	-0.04	0.11	0.96	0.747	Kalenjin	1.41	0.3	4.11	< 0.001
40-44	-0.08	0.12	0.92	0.481	Kamba	0.97	0.3	2.64	0.001
45-49	-0.39	0.14	0.68	0.005	Kikuyu	-0.16	0.29	0.85	0.587
					Kisii	1.42	0.31	4.15	< 0.001
Region					Luhya	1.4	0.29	4.04	< 0.001
Nairobi					Luo	1.02	0.3	2.77	< 0.001
Central	1.8	0.14	6.08	< 0.001	Masai	1.94	0.36	6.98	< 0.001
Coast	1.39	0.15	4.03	< 0.001	Meru	-0.22	0.32	0.81	0.496
Eastern	2.08	0.14	7.99	< 0.001	Mijikenda	1.57	0.32	4.8	< 0.001
Nyanza	2.15	0.15	8.62	< 0.001	Somali	3.1	0.36	22.28	< 0.001
Rift Valley	1.73	0.14	5.62	< 0.001	Taita/Taveta	0.3	0.4	1.35	0.445336
Western	1.72	0.15	5.56	< 0.001	Turkana	2.99	0.43	19.86	< 0.001
North Eastern	2.86	0.21	17.46	< 0.001	Kuria	2.75	0.51	15.7	< 0.001

4. Discussion

Asset based measures of poverty are increasingly being used there are some limitation on their use.

- The asset- based measures are more reflective of the long-run household wealth, failing to capture short-run wealth to the household (Filmer and Pritchett 2001). Therefore, if the outcome of interest is associated with current resources to the household, then an index based on asset may not be the best measure
- The second issue is that the ownership does not capture the quality of the asset, (Falkingham and Namazie (2002)).
- Some variables may have a different relationship with the asset index across sub-groups; for example ownership of farm land may be more reflective of wealth in rural areas than urban areas.

The multi-variate analysis shows that increases in educational attainment have an important impact on reducing the probability that a household is poor. The logistic model shows that a rural family has a high probability of being poor. The rural/urban variable is statistically significant and this variable increases the odds of a household being poor significantly. The other demographic factors that increase the probability of being poor are the age of the household head, religion, region and ethnicity. Size of household when tested as a univariate model, has statistically significant with the social economic status but it's not significant when included in the multivariate analysis. The rural/urban variability can be argued that the assets included in the asset index are by their nature urban rather than rural and therefore are biased against rural areas. Indeed in most of African countries, the governments regard the provision of formal housing, water and sanitation services as naturally urban services, but as the countries develop it would not be amiss for the rural population to strive towards having piped water, flush toilets and good housing characteristics. It is possible that important changes may take place in the economic situation of many households, but the asset indices may remain unchanged. That being the case then we cannot use the asset index to measure short or medium term social welfare of a household.

References

[1] Basic Report on well-being in Kenya (2005/06) Annals of Kenya Integrated Household Budget Survey.

[2] Booyen, R (2002). Using Demographic and Health Surveys To Measure poverty. An Application to South Africa

[3] Collett, D (2003). *Modeling Binary Data*. Chapman and Hall/CRC texts in statistical science series.

[4] Central Bureau of statistics (Kenya),(2004), *Kenya Demographic and Health Survey* (2003) *Report*.

[5] Filmer, D (2000).The Structure of Social Disparities in Education: Gender and Wealth. *Annals of World Bank Policy Research working paper 2268*.

[6] Filmer, D. and Pritchett, L (2001). Estimating wealth effects without expenditure Data or Tear: Educational enrolment in India. *Annals of Mathematical Statistics*, **14**, 436-440.

[7] Filmer, D. and Pritchett, L (1999). The Effect of Household Wealth on Education Attainment: Evidence from 35 Countries. *Population and Development Review*, Vol. 25, 1, 85-120.

[8] Foster,J., Greer, J and Thorbecke, E (1984). A class of decomposable poverty measures. *Econometrica*, **52**, 761-765.

[9] Houweling TAJ, Kunst AE, Mackenbach JP. 2003. Measuring health inequality among children in developing countries: does the choice of the indicator of economic status matter? *International Journal for Equity in Health* **2**, 8.

[10] *Kyereme*, S., Thorbecke, E. (1991). Factors affecting *food* poverty in Ghana. *Journal of Development Studies*, **28**, 39-52.

[11] Rodriguez, A.G., and S.M. Smith (1994). *A* comparison of determinants of urban, rural *and* farm poverty in Costa Rica. *World Development*, **22** (3), 381-397

[12] Mwabu, Germano, Wafula Masai, Rachel Gesami, Jane Kirimi, Godfrey Ndeng'e, Tabitha Kiriti, Francis Munene, Margaret Chemngich and Jane Mariara, 2000, *Poverty in Kenya: Profile and Determinant*', Nairobi: University of Nairobi and Ministry of Finance and Planning.

[13] Oyugi, Lineth Nyaboke (2000). The determinants of poverty in Kenya. MA Thesis, Economic Department. *University of Nairobi*

[14] Sahn, D.and Stifel, D (2003). Exploring Alternative measure of welfare in Absence of Expenditure Data. *Review of income and wealth*, **49**: 463-489

[15] Sahn, D.and Stifel, D (2000). Poverty Comparisons over time and across countries in Africa. *World Development*, **28**: 2123-2155

[16] Snijder and Bosker (1999). Multilevel Analysis: An introduction to basic and advanced multilevel modeling. Sage publications Ltd

[17] Vyas, S and Kumaranayake, L (2006). Constructing social-economic status indices, How to use PCA. *Health Policy and Planning*, **21**, 6, 459-468

[18] World Bank (2000). *Development Report 2000/01, attacking poverty*. Washington DC World Bank.