NEW DEVELOPMENTS IN PROXY MEANS TEST ANALYSIS[1] USING DATA FROM ETHIOPIA

VENANZIO VELLA[§]

MAURIZIO VICHI [*]

AUGUST 6, 1999

§ The World Bank, AFTH4, Vvella@worldbank.org

* Maurizio Vichi, professor of statistics at "University G. D'Annunzio" Dpt. Metodi Quantitativi, Viale Pindaro 42, I-65127 Pescara, Italy " em: vichi@dmqte.unich.it.

[1] The study was financed by the Thematic Group on Inequality and Poverty and was carried out in collaboration with Prof. Maurizio Vichi, who was financed by the Italian Trust Fund. Prof. Maurizio Vichi is author of several publications on international journals. He is Professor of Quantitative Methods at the University of Chieti, Faculty of economics, and he is the Secretary General of the Italian Statistical Society.

Abstract

This analysis was carried out on the 1995/96 Ethiopia Household Budget Survey to achieve two separate but complementary objectives. The first one was to test the discriminatory power of Ordinary Least Square (OLS), Discriminant Analysis (DA) and Logistic regression (LR) in identifying the households below and above the 40th percentile of expenditures. The second objective was to capture the multidimensional aspects of poverty through the Non Linear Component Analysis and Cluster Analysis.

The OLS produced an overall correct classification (hit ratio) of 62.3 percent with 54 percent of the poor correctly identified. When the outliers[2] were excluded from the analysis, the hit ratio rose to 62.6 percent, with 55.7 percent of the poor correctly classified. The scarce increment of the hit ratio suggests that the violation of some of the regression assumptions do not influence the results of the OLS, indicating its robustness as a technique for proxy means testing.

The proxy means test based on the DA improved the hit ratio to 62.8 percent, with a much better identification of the poor (68.9%). Furthermore, the DA estimated households' probabilities to belong to the eligible and non-eligible group. This allowed to assess the effect that the selection of given levels of probability of belonging or not to the target group had on the accuracy of the test. For example, selecting a probability of not less than .52 (2% greater than the chance) of correctly identifying the target group, produced a small proportion of unclassified households (12.3%) which belonged to the interval between 50 and 52 percent. But the hit ratio increased to 65 percent and the proportion of poor correctly identified increased to 71 percent for the remaining sample. This flexible use of the test could help to adjust undercoverage or leakage according to the need of a given program. For example when the probability was fixed to reduce undercoverage, 23 percent of the households were excluded from the classification, while for the remaining sample, the hit ratio was the same as that obtained through the OLS (62%) but the proportion of poor households correctly classified increased to 85 percent.

[2] +- 3 standard deviations of the studentized residuals

LR was more efficient in terms of hit ratio (66.1%), with a high proportion (76.2%) of non poor households correctly classified and a low proportion (52.4) of poor correctly classified. Also in the case of LR, different probabilities to belong to the poor and non-poor groups could be used to reduce leakage and/ or undercoverage.

.

To achieve the second objective, the NLPCA was used to optimally scale the nominal, ordinal and continuous proxies of welfare, while the CA was applied on the optimally scaled variables to partition the households in five clusters characterized by different levels of welfare. The prevalence in welfare characteristics and malnutrition across clusters was used to validate if this technique had successfully identified group of disadvantages household.

The prevalence of proxies of deprivation such as illiteracy, use of unprotected water supplies between the first and fifth cluster was higher than between the first and fifth quintiles of expenditures. This suggests the classification provided by the CA was more powerful in forming groups characterized by different living conditions compared with the groups formed according to the expenditure quintiles. This was confirmed by the high variation in stunting prevalence, which was 44 percent in the worst off cluster and 20 percent in the better off cluster. While wasting was 9 and 4 percent respectively in the in the worst off and the better off cluster. The variation in prevalence of malnutrition was much lower across the expenditure quintiles; suggesting that NLPCA and CA were more powerful in capturing the multidimensional aspects of poverty.

The promising results of this study justify a wider application of these analytical techniques to available data sets to improve the accuracy and flexibility of proxy means and to use pilot test their applicability in operational settings.

Keywords: Proxy means test, OLS regression, Discriminant Analysis, Logistic Regression, Cluster Analysis, Nonlinear Principal Component Analysis

1. Introduction

Testing is used to screen individuals, households, geographical areas and other units for targeting purposes. Tests are used in health, social, economic and other programs to cover as much as possible those who need the benefits and exclude those who do not. The test must be validated against a standard, which at the best of out knowledge, reflect the characteristics of the target group.

It is important to clarify the objective of the testing and the standard against which the test is validated. For example, in a public health program against hypertension, the standard is usually the blood pressure above a certain cut off. Similarly, in a supplementary feeding program the standard can be provided by an anthropometric cut off point. In a poverty alleviation program the standard can be based on a money metric (i.e. expenditures).

Because applying the above mentioned tests on large population can be too costly, proxy testing is commonly used. In the case of proxy means testing, welfare proxies (i.e. household characteristics) are identified and are provided a weight according to their correlation with a higher level standard measure (i.e. expenditure). These weights will then be applied to household to decide if their total score make them eligible to receive certain benefits. The accuracy of the classification is then validated against the higher level standard, which in the money metric is usually expenditure below and above the 40^{th} percentile.

This paper does not aim to summarize the substantial literature on proxy means testing but to quote some key references. Ravallion et al showed that macro indicators are not very promising for targeting geographic areas in India (Datt and Ravallion 1993) and Indonesia (Ravallion 1993). Hentschel at al (1998) were more successful in identifying macro targeting approaches through modeling consumption from a high quality survey in Ecuador, using predictors which were also collected in the census. Haddad et al. (1991) analyzed several surveys to find a few household variables to be used as proxies for food security. Glewwe and Kanaan (1989) utilized regression

analysis to select predictors based on household characteristics. Grosh R. M. and Baker J.L. (1995) conducted simulations ion the application of household variables for proxy means testing and they found that even imperfect targeting could cut down leakage substantially.

Whatever statistical procedures are used to build a proxy means test, it is essential to define precisely the target group (Grosh & Glinskaya 1998), according to the objective of the program. If we use the money metric, the eligible group may be defined according to a given poverty line such as the 40^{th} expenditure percentile. The overall accuracy of the test is provided by the overall proportion of eligible and ineligible households classified correctly (hit ratio). The ability of a test to classify correctly a high proportion of eligible is called sensitivity, while the ability to classify correctly a high proportion of ineligible is called specificity. Lower cut off points to define eligibility are accompanied with higher sensitivity and lower specificity with subsequent higher under-coverage but lower leakage. The increase in the cut off point is accompanied by the opposite situation. Because sensitivity ands specifity are inversely correlated they can be used to meet the specific objective of a program to avoid under-coverage or leakage according to a more generous or more conservative cut off point to define eligibility.

Because the proxy means tests are based on their correlation with a higher level standard (i.e. the money metric), the validity of the tests in identifying the eligible group is as good as the standard used. For example a high sensitivity test based on the money metric may achieve the objective of identify a high proportion of those below the 40^{th} percentile of expenditure but it may be failing to reach the most needy if expenditures are not appropriate to capture deprived household. The issue is therefore to clarify if in certain countries expenditures are indeed the golden standard to capture the multidimensional aspects of poverty.

Even in those countries where expenditures are appropriate, they can be affected by several problems. These include inaccurate reporting due to recall or other bias, and the exclusion of the imputed value of household, durable goods and home-grown-crop (Grosh & Gliskaya 1998). The cut off point may be arbitrarily selected (i.e. 40^{th}

percentile) with subsequent failure to define a more objective criteria related to a minimum expenditure to cover essential needs. Furthermore, expenditures do not capture the whole dimension of poverty, which is as difficult o measure as the concept of human dignity (Singer 1983) and as mentioned by the 1997 UNDP Human Development Report "*Poverty must be addressed in all its dimensions, not income alone*. [(3)]". Sophisticated statistical techniques such as *Nonlinear Principal Component Analysis and CA can help to make a dent on the complex aspect of defining and measuring poverty according to a holistic metric.*

This study had therefore two different but complementary objectives: (a) improve the means testing techniques based on the money metric and (b) to define poverty according to other criteria. To achieve the first objective, the following analytical techniques were applied: Ordinary Least-Squares (OLS) as per the original proxy means-test analysis proposed by Grosh & Baker (1995), Grosh & Glinskaya, (1998); *Discriminant Analysis* (DA) and *Logistic Regression* (LR). DA and LR, are alternative and appropriate methods to find discrimination criteria to predict household group membership. Cross-validation will be used to compare the accuracy of the prediction of the means tests based on these statistical techniques.

To achieve the second objective, NLPCA was used to construct dimensions of poverty based on welfare proxies and CA clusters households with similar scores for these dimensions. Because of the different metric used, the results of the NLPCA and CA were validated against stunting and wasting, which are well known biological standards of deprivation.

2. Methodology and data sources

The data were from the 1995/96 Ethiopia Household Income and Budget Expenditure Survey, which was financed by the World Bank. The independent variables were selected taking into consideration that irrelevant variables can reduce model parsimony, mask or replace the effects of more useful variables; and that the omission of relevant variables

([3]) *Human Development Report*, United Nations Development Programme, Oxford University Press, 1997.

can negatively affect the predictive accuracy of the analysis. The selected variables were those commonly used in proxy means-test mechanism, including:

- characteristics of the head of the household such as age, gender, and education ;
- ownership of assets such as land, farming, animals, house, vehicles, bicycles, motorcycles, radios, TVs, refrigerators, electric stoves, sickles, tractors;
- household characteristics such as dependency ratio, sources of water, cooking fuel and sanitation facilities; and
- access to services such as health centers, post offices and primary schools.

Because regression and other multivariate analyses are seriously effected by missing data, only data with no missing data were used ([4]). Of the total sample of 11687 households, 8823 with no missing data were selected for the analysis. This is unlikely to have biased the analysis, as the sample without missing data did not differ significantly from the sample with missing data in terms of expenditures and other welfare variables.

OLS *regression*, *Discriminant Analysis* and *Logistic Regression* (see Annex) were applied to find the correct discriminating criterion (test) to predict eligibility or ineligibility of household. The procedure to validate the proxy means test was the *cross-validation* (Green & Carroll 1978) which involves the random assignment of the households to the analysis and holdout samples. The analysis sample was used to construct the proxy mean test function while the holdout sample was used to validate it. Cross-validation avoids the "overfitting" of the regression or discriminant or logistic regression analyses by allowing their validation on a totally separate sample. There are no definitive guidelines to divide the sample into the analysis and holdout groups. We utilized the 60 - 40 percent split between the analysis and the holdout samples according to the method used by several authors (Green & Carroll 1978; Hubert, Wisenbaker & Smith 1987; Hosmer & Lemeshow 1989). The 8823 households were randomly assigned to the analyses (5297) and holdout sample (3526).

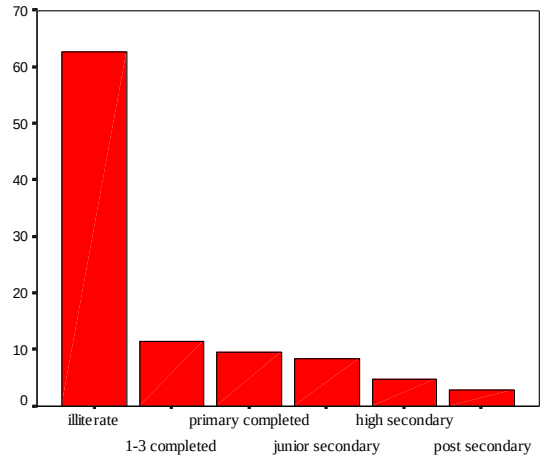
The frequency distribution of expenditures (Table 1) and welfare variables (Fig 1a-1h) show a low variability, suggesting that some of the variables were not ideal to

([4]) Cases with missing data Can produce indefinite correlation matrices with subsequent incorrect results such as R^2 greater than one in regression analysis or negative variance in principal component analysis.

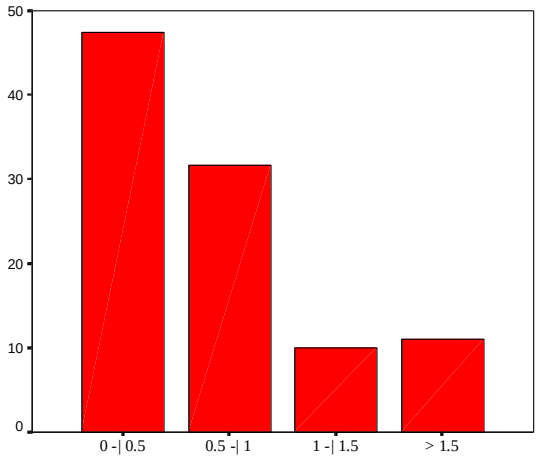
represent welfare characteristics in Ethiopia. The households were characterized by a high illiteracy rate; low possession of items such as TVs, radios, motor-vehicles, fridge; use of unprotected water and cheep cooking fuel (i.e. dung).

Figures 1a-1h: distributions of some variables associated with welfare

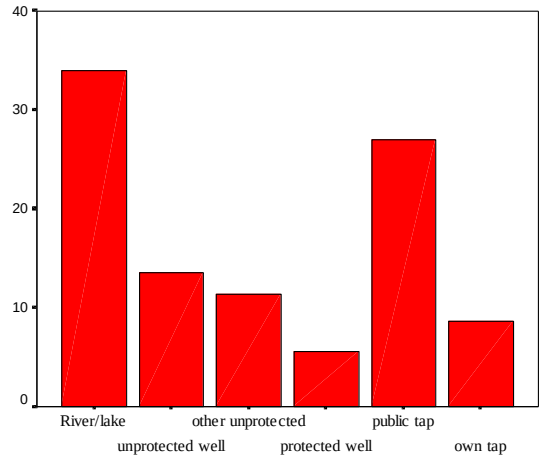
Education of the head of the household



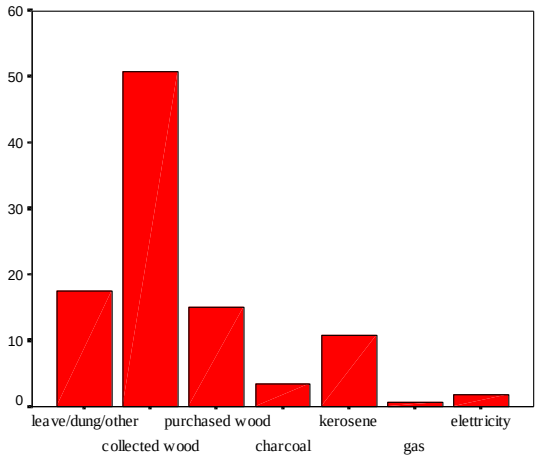
Dependency ratio



Source of water used



Source of cooking fuel



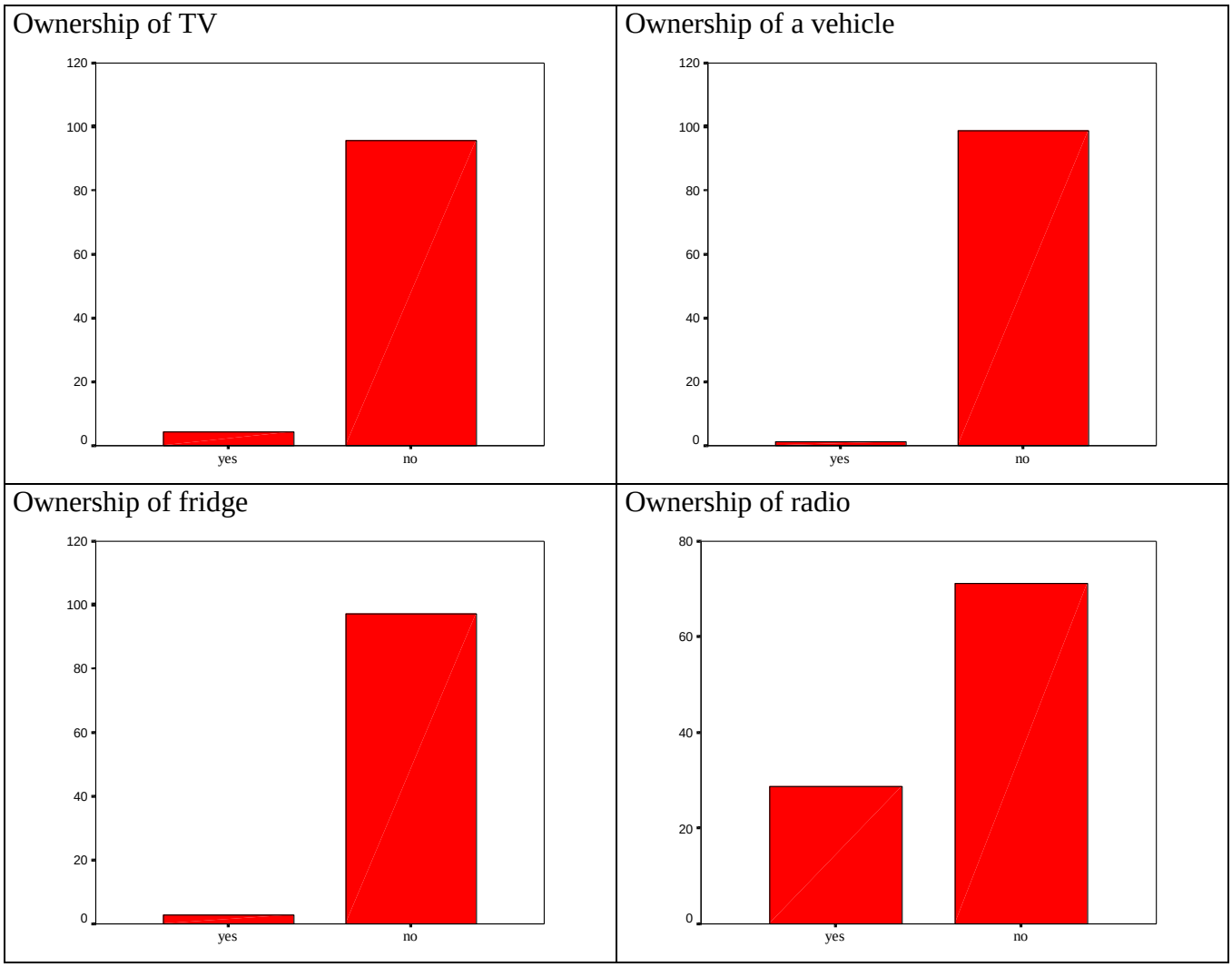


Table 1: Per capita expenditure quintiles (Ethiopian birr)

Percent	Expenditure
20	697
40	946
60	1260
80	1780

3 Results

Sections 3.1 through 3.3 deal with the results of the proxy mean testing obtained through OLS, DA and R; and section 3.4 deals with the definition of poverty obtained through the NLPCA and CA.

3.1 Proxy Means-Test by Ordinary Least-Squares Regression

Taking into account the original proxy means-test proposed by Grosh & Baker (1995), information on variables other than consumption were used to predict the total per capita expenditure. The prediction is obtained through Ordinary Least Squares (OLS) regression and the household is considered poor if the predicted value of the expenditure is below the ([5]) 40th percentile (Grosh & Glinkava, 1998).

As already noted by several authors, since the independent variables used in OLS are frequently correlated between each other and most of them are categorical, the OLS regression is not the best choice, because the assumptions of regression analysis, such as normality and multicollinearity ([6]) are violated. However, since the purpose of the analysis is neither to interpret the regression coefficients nor to predict welfare, but to predict eligibility of the household, the good results obtained in previous analyses (Grosh & Baker 1995) suggest to continue to explore ways to improve accuracy of the prediction through this procedure. Furthermore, we will show ways to reduce the violation of the basic regression assumptions and how these violations did not affect the results of the analysis.

The results of the stepwise approach used to finalize a parsimonious model with the best fit based on the per capita total expenditure as dependent variable are reported in Tables 2 and 5.

([5]) Also the 40th percentile of the observed distribution could be used.

([6]) Collinearity is the association, measured through correlation, between two independent variables. Multicollinearity refers to the correlation among three or more independent variables. Although there is a distinction in the two terms often they are used interchangeably.

Table 2: Variables included in the regression model

Variables included in the regression model	Coefficient	Standard Error	Standardized coefficients	t	Sig.
Constant	2916.696	226.426		12.881	.000
Source of cooking fuel	78.656	11.649	.088	6.752	.000
Dung/other	-152.537	480.228	-.055	-.318	.751
Collected wood	-58.964	479.935	-.028	-.123	.902
Purchased wood	9.441	480.342	.003	.020	.984
Charcoal	70.072	482.296	.012	.145	.884
Kerosene	111.914	480.711	.033	.233	.816
Gas	479.673	495.251	.035	.969	.333
Electricity and other	-77.014	484.763	-.010	-.159	.874
Dependency ratio	-132.249	10.125	-.133	-13.062	.000
0 -| 0.5	-82.093	67.260	-.039	-1.221	.222
0.5 -| 1.0	-394.663	67.338	-.172	-5.861	.000
1.0 -| 1.5	-477.964	71.969	-.134	-6.641	.000
> 1.5	-454.320	69.016	-.133	-6.583	.000
Education head of the household	100.958	10.415	.124	9.694	.000
*illiterate	-	-	-	-	-
1 - 3 completed	33.338	29.593	.010	1.127	.260
primary completed	70.846	33.810	.020	2.095	.036
junior secondary	201.551	37.416	.053	5.387	.000
high secondary	553.600	47.949	.113	11.546	.000
post secondary	906.958	60.320	.144	15.036	.000
Ownership of radio - no	-209.027	31.059	-.087	-6.730	.000
Use of the primary school - no	155.725	22.103	.076	7.045	.000
Gender of the head of the household - female	142.827	24.601	.060	5.806	.000
Ownership of TV set - no	-279.118	65.683	-.052	-4.249	.000
Age of the head of the household	-78.874	14.263	-.059	-5.530	.000
Age 0 -| 15	-71.370	319.728	-.002	-.223	.823
Age 16 -| 25	409.898	35.217	.105	11.639	.000
Age 26 -| 35	201.916	22.640	.082	8.918	.000
*Age 36 -| 65	-	-	-	-	-
Age > 65	187.198	34.562	.052	5.416	.000
Use of the health center - no	-137.012	31.087	-.044	-4.407	.000
Ownership of a vehicle - no	-316.494	101.597	-.035	-3.115	.002
Toilet facilities	-46.106	15.028	-.041	-3.068	.002
Flush toilet	204.462	99.362	.027	2.058	.040
pit	-105.834	74.285	-.046	-1.425	.154
Container (form hh items)	550.521	201.208	.025	2.736	.006
other	-162.812	70.612	-.073	-2.306	.021

*Category excluded by the model

Table 3: Variables excluded from the regression model

Variables excluded from the model After 11 steps	Standardized Coefficient In	t	Sig.	Partial Correlation	Statistics Collinearity
Ownership of transport animals	.021	1.846	.065	.020	.796
Number of buildings owned	.007	.646	.518	.007	.969
Source of water	-.010	-.723	.470	-.008	.558
Ownership of house	.008	.814	.415	.009	.972
Ownership of land	.006	.578	.563	.006	.987
Ownership of refrigerator	-.007	-.553	.580	-.006	.584
Ownership of electric stove	-.012	-.960	.337	-.010	.637
Ownership of sickle	-.009	-.789	.430	-.008	.812
Ownership of tractor	.000	-.006	.995	.000	.999
Ownership of farming animals	-.015	-1.367	.172	-.015	.812
Ownership of bicycle	.019	1.864	.062	.020	.911
Ownership of motorcycle	.012	1.133	.257	.012	.941
Use of post office	-.014	-1.217	.223	-.013	.764

Table 4: Fit of the model

R	R-squared	R-squared corrected	Standard error of estimate
.359	.129	.128	930.4773
With dummy variables			
.426	.182	.179	958.0064

Table 5: Analysis of Variance

	Sum of squares	df	Mean square error	F ratio	Sig.
Regression	1131729296.212	11	102884481.474	118.833	.000
Residual	7629323674.865	5286	865787.979		

Figure 2 shows the plot involving the residuals versus the predicted dependent variable. The plot confirms that assumptions required by the regression model are not strictly met because many residuals are not falling randomly, with relatively equal dispersion around zero, but they are located in the positive part of the plot. These large positive residuals reduce the fit of the model and should be removed from the analysis.

Figure 3 shows the normal probability plot, where the standardized residuals are compared with the normal distribution. If the distribution of residuals is normal, the plotted line follows closely the diagonal. The plot indicates an s-shaped curve showing that the distribution of residuals is not strictly normal.

Figure 2: Standardized residuals versus predicted dep. Var.

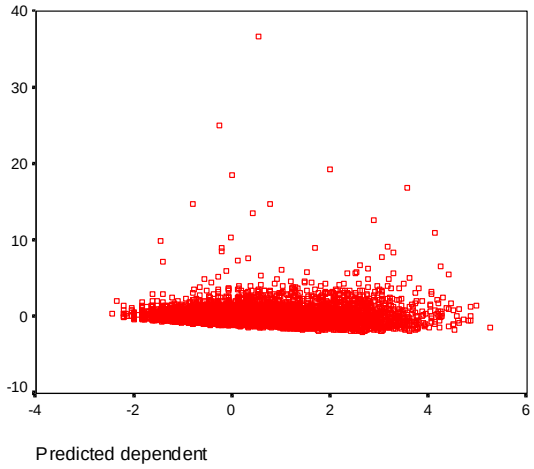
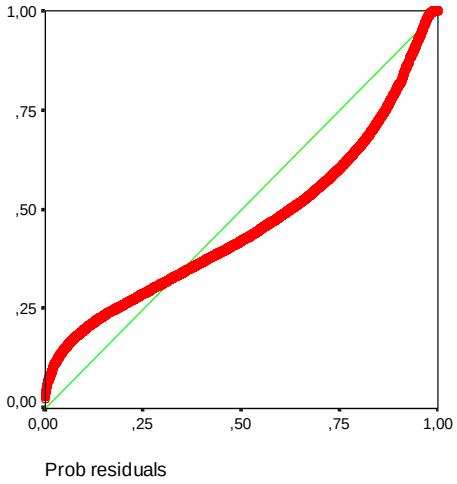


Figure 3: Normal probability plot



The value of the 40th percentile predicted expenditure was 1199; thus, the households with per capita total expenditure below or equal to 1199 were considered eligible.

Prediction of eligibility is subject to two types of error. Rejection of the null hypothesis when, in fact, it is true; or acceptance of the null hypothesis when it is false; also known as type I and II, or exclusion and inclusion errors. Table 6 reports the classification matrix with the percentage of correctly classified households computed on the holdout (validation) sample.

Table 6: Classification matrix obtained by proxy means-test based on OLS regression

Observed	Predicted			percent Correctly Classified
	Eligible	Ineligible	Total	
Poor	817	670	1487	54.9%
Non poor	659	1380	2039	67.7%
Total	1476	2050	3526	62.3%

The systematic presence of observations in the positive part of the plot of residuals (Fig 2) has a disproportionate effect on the regression results. These observations may be due to (a) errors in data collection and/or data entry; (b) a valid but exceptional observation that is explained by an extraordinary situation; (c) an ordinary observation in its individual characteristics but exceptional in its combination of characteristics. When the correction error is not possible, it is better to delete the case. With valid, but exceptional observations, deletion of the case is warranted unless variables reflecting the extraordinary situation are included in the regression equation. In all situations, the researcher is encouraged to delete only truly exceptional observations, but still guard against deleting observations that, while different, are representative of the population.

The diagnostic tool to identify influential observations involves the identification of outliers, i.e., observations not predicted well by the regression equation that have large residuals. The statistical significant large residuals were those corresponding to the studentized residuals larger than $t=\pm 3$, corresponding to less than one percent population (75 observations). The deletion of these outliers, improved the fit of the model which had two extra variable compared to the previous model (Tables 7 and 8). The plot of residuals (Figure 4-5) of the last model were more in line with the basic assumptions of regression analysis, although the normal probability plot still showed a non normal distribution of the residuals.

However, even if the fit of the regression model improved (R^2=0.201), the

predictive power of the eligibility test did not improved much (Table 11) and a further deletion of outliers (t==±2) did not significantly improve accuracy of the test. The fact that improvement in the violation of the basic assumptions of OLS did not influence the predicting power, suggests that OLS is not affected by minor violation of regression assumption and it is a suitable technique for proxy means-test.

Table 7: Variables included in the regression after deleting outliers (observations outside ± 3SD)

Variables included in the regression model	coefficient	Standard Error	Standardized coefficients	t	Sig.
Constant	2459.094	154.528		15.914	.000
Education head of the household	94.339	7.085	.165	13.315	.000
*illiterate	-	-	-	-	-
1 - 3 completed	62.151	17.927	.030	3.467	.001
primary completed	101.904	20.611	.046	4.944	.000
junior secondary	177.178	22.916	.074	7.732	.000
high secondary	468.866	29.685	.151	15.795	.000
post secondary	717.377	37.877	.176	18.940	.000
Dependency ratio	-126.934	6.806	-.184	-18.650	.000
*0 -| 0.5	-	-	-	-	-
0.5 -| 1.0	-234.991	12.808	-.166	-18.347	.000
1.0 -| 1.5	-319.577	19.312	-.146	-16.548	.000
> 1.5	-313.600	18.915	-.149	-16.579	.000
Ownership of radio - no	-225.945	20.904	-.135	-10.809	.000
Source of cooking fuel	63.652	7.944	.102	8.013	.000
Leave/dung/other	-64.682	15.234	-.037	-4.246	.000
*Collected wood	-	-	-	-	-
Purchased wood	40.752	19.682	.022	2.070	.038
charcoal	127.748	33.809	.034	3.779	.000
kerosene	27.495	24.167	.013	1.138	.255
gas	455.142	78.438	.052	5.803	.000
Electricity and other	-42.937	45.620	-.009	-.941	.347
Use of the primary school - no	125.038	14.934	.088	8.373	.000
Gender of the head of the household - male	137.850	16.989	.083	8.114	.000
Ownership of TV set - no	-198.980	44.358	-.053	-4.486	.000
Use of the health center - no	-111.003	20.866	-.052	-5.320	.000
Age of the head of the household	-38.396	9.580	-.041	-4.008	.000
Age 0 -| 15	-57.833	191.464	-.002	-.302	.763
Age 16 -| 25	242.613	21.655	.098	11.204	.000
Age 26 -| 35	115.818	13.779	.075	8.405	.000
*Age 36 -| 65	-	-	-	-	-
Age > 65	179.098	21.257	.073	8.426	.000
Ownership of transport animals - no	55.097	14.888	.040	3.701	.000
Ownership of a vehicle - no	-224.167	68.835	-.035	-3.257	.001
Use of post office - no	-45.781	16.815	-.030	-2.723	.006
Toilet facilities	-27.545	10.229	-.035	-2.693	.007
Flush toilet	32.885	61.090	.007	.538	.590
Pit	-117.137	45.164	-.081	-2.594	.010
Container	3.988	122.778	.000	.032	.974
*Field/forest	-	-	-	-	-
Other	-133.982	42.841	-.096	-3.127	.002

Table 8: Variables excluded from the regression model after deleting outliers (observations outside ± 3 SD)

Variables excluded from the model After 11 steps	Standardized Coefficient In	t	Sig.	Partial Correlation	Statistics Collinearity
Number of buildings owned	.010	1.050	.294	.011	.963
Source of water	.010	.755	.450	.008	.548
Ownership of house	.004	.417	.677	.004	.973
Ownership of land	-.009	-.948	.343	-.010	.982
Ownership of refrigerator	-.012	-.996	.319	-.011	.593
Ownership of electric stove	-.011	-.922	.356	-.010	.639
Ownership of sickle	-.004	-.324	.746	-.003	.743
Ownership of tractor	-.001	-.083	.934	-.001	.999
Ownership of farming animals	.024	.864	.388	.009	.122
Ownership of bicycle	.008	.824	.410	.009	.909
Ownership of motorcycle	.009	.941	.347	.010	.939

Table 9: Fit of regression model, after deleting outliers (observations outside ± 3SD)

R	R-squared	R-squared corrected	Standard error of estimate
.448	.201	.200	620.3392
With dummy variables			
.501	.251	.249	573.6634

Table 10: Analysis of Variance

	Sum of squares	df	Mean square error	F ratio	Sig.
Regression	844705835.142	13	64977371.934	168.851	.000
Residual	3361024120.000	5209	384820.714		

Figure 4: Standardized residuals versus predicted dep. Var. after deleting outliers (observations outside ± 3SD)

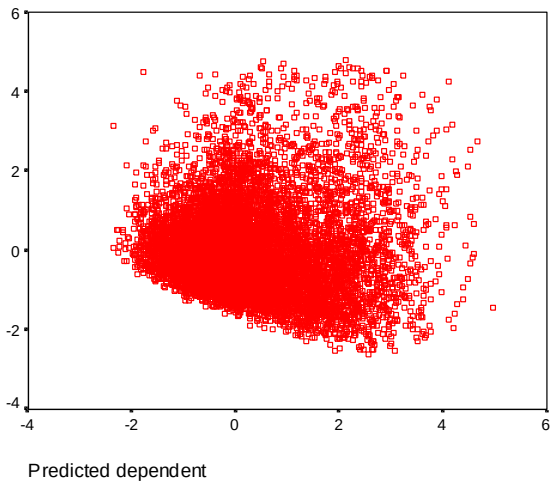


Figure 5: Normal probability plot after deleting outliers (observations outside ± 3SD)

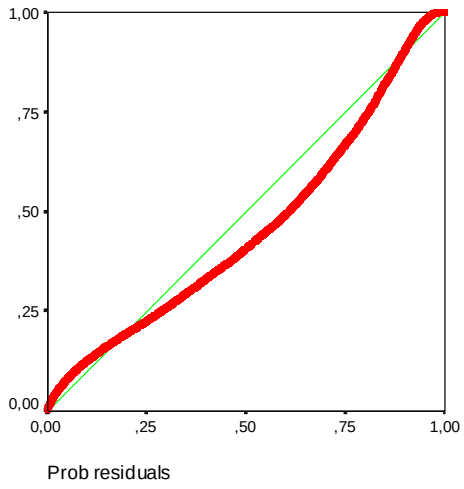


Table 11: Classification matrix obtained by proxy means-test based on OLS after deleting outliers *

Observed	Predicted			% Correctly Classified
	Eligible	Ineligible	Total	
Poor	829	658	1487	55.7%
Non poor	659	1380	2039	67.7%
Total	1488	2038	3526	62.6%

*Observations outside ± 3SD

When dummy variables were included, the R^2 increased from .18 to .25 but the predictive power did not increase, suggesting that multicollinearity did not affect the results of the OLS. This is based on the assumption that multicollinearity limits the size of the coefficient of determination R^2 ([7]), and the ability of the regression procedure to represent the contribution of each independent variable in explaining the regression variate. A decrease of collinearity is usually accompanied by an increase in total

([7]) If the correlation between X1 and Y were 0.6, a model with X1 would explain 36 percent of the variance of Y. If correlation between X2 and Y were 0.5, X2 would explain 25 percent of the variance of Y. If X1 and X2 were not correlated between each other (independent), a model with X1 and X2 would explain 61

variance of Y explained by the regression model, which should be accompanied by an improved discriminating power on the dependent. The fact that the increased R^2 did not change the predictivity of the test suggests that multicollinearity did not influence the results of the OLS.

3.2. Proxy Means-Test by Discriminant Analysis

In the previous section regression analysis was used to estimate welfare on a basis of a set o variables correlated with it and easy to collect. However, the household estimation of welfare is not the real objective of the proxy means-test. The proxy means-test involves the construction of a discriminant tool, to decide if a household belongs or not to the target group. To achieve this objective, Discriminant Analysis (DA) was applied on the analysis sample to identify the discriminant function to be tested on the holdout sample. For comparability sake, DA was applied on the same variables used in the OSL.

A sequential estimation method, similar to step-wise regression, was used to identify subsets of variables with the greatest discriminatory power. The stepwise approach began by choosing the single best discriminating variable which was then paired with each one of the other independent variables. The variable was entered if it improved the discriminating power of the model. As additional variables were included, some previously selected variables were removed if they did not contribute significantly to the model. The final model is shown in Tables 12 and 13.

---

percent of the variation in Y. If X1 and X2 were correlated between each other, the total variance of Y explained by the regression model decrease.

Table 12: Variables in the final model obtained through DA

Variables in the discrimination model after eleven steps.	Tolerance	F to remove	Wilks' Lambda
Ownership of radio	.654	116.132	.907
Dependency ratio	.952	238.681	.919
Education head household	.644	68.520	.902
Use of the health center	.970	35.314	.899
Ownership of transport animals	.797	38.347	.899
Gender of the head of household	.877	26.369	.898
Source of cooking fuel	.650	29.298	.898
Use of primary school	.834	20.255	.897
Number of buildings owned	.967	16.448	.897
Age of head of household	.890	16.294	.897
Use the post office	.792	5.139	.896

Table 13: Variables excluded

Variables not in the model after eleven steps	Tolerance	Min Tolerance	F to enter	Wilks' Lambda
Source of water	.602	.602	.315	.895
Toilet facility	.547	.547	.268	.895
Ownership of house	.972	.642	2.899	.895
Ownership of land	.982	.643	.525	.895
Ownership of vehicle	.940	.638	.531	.895
Ownership of TV set	.786	.614	1.488	.895
Ownership of refrigerator	.871	.630	1.441	.895
Ownership of electric stove	.699	.613	.487	.895
Ownership of sickle	.748	.627	1.283	.895
Ownership of tractor	.999	.643	.279	.895
Ownership of farming animals	.122	.120	.006	.895
Ownership of bicycle	.954	.640	.748	.895
Ownership of motorcycle	.994	.643	.096	.895

The tolerance (in Table 12 and 13) is the proportion of the variation of the independent variable not explained by the variables already in the discriminant function. A tolerance of 0 means that the independent variable under consideration is a perfect linear combination of independent variables already in the discriminant function, hence showing multicollinearity. When tolerance is one, an independent variable is totally independent of other variables already in the model.

The level of significance of the DA model (Table 14) was measured through Wilks' lambda and the partial *F* ratio. Large F values indicate greater discriminatory power. Lambda varies from 0 to 1, with 0 meaning that group means differ (thus the more the variable differentiates the groups), and 1 meaning that group means are the same.

Table 14: Significance testing of group differences after step eleven

Lambda	df1	df2	df3	Partial F ratio			
				Statistic	df1	df2	Sig.
.895	11	1	8821	93.983	11	8811.000	.000

Df = degree of freedom

The multivariate aspects of the model are illustrated under the heading *Canonical Discriminant function* (Table 15). The discriminant function is highly significant and displays a canonical correlation 0.324. Its square (0.10) can be interpreted as the variance in the dependent variable accounted for by this model. Although the value was low, it was sufficient to show a discriminating power superior to the OLS model, as can be seen from Table 11[ii].

Table 15: Canonical discriminant function coefficient

Eigenvalue	% of variance	% cumulative	Canonical Correlation	After function	Wilks' Lambda	Chi-square	df	Sig.
				0	.895	978.027	11	.000
.117	100.0	100.0	.324					

Table 16 and 17 show the canonical discriminant function coefficients and the discriminant loadings. The coefficients are the variables' weight defining the discriminant function, while the discriminant loadings measure the linear correlation between each independent variable and the discriminant function. Higher the loadings, greater the variance that the independent variable share with the discriminant function. Loadings are ranked in magnitude order, independently from the sign, to indicate which variables are substantive discriminators worthy of note.

The classification functions (Table 18), also known as Fisher's linear discriminant functions, one for each group (poor) and (non-poor), can be used to classify new households as belongings to one of the two groups. The values of the independent variables for the new household are inserted in the two classification functions and a classification score for each group is calculated for that observation. The observation is then classified into the group with the highest classification score.

A measure of success of discriminant analysis is its ability to define a disciminant function that results in significant different centroids (Table 19). These corresponds to the average profiles of the two groups and the difference between centroids can be measured in terms of Mahalanobis D^2 measure to test for statistically significant differences.

Cross-validation approach was used to assess the accuracy of the model in identification eligible and ineligible households. The discriminant function, applied on the holdout sample, produced a higher hit ratio (Table 11[ii]) than one produced through the OLS function. These results suggest that discriminant analysis is a more appropriate technique to predict eligible households.[8]

[8] No significant changes in the classification accuracy were observed, when outliers were excluded.

Table 16: Canonical discriminant function coefficients

Independent variable	Standardized	Unstandardized
Age head of the household	.141	.189
Ownership transport animals	-.228	-.457
Dependency ratio	.514	.520
Gender of the head of the household	-.180	-.431
Education of the head of the household	-.338	-.282
Number of buildings owned	-.135	-.264
Source of cooking fuel	-.220	-.200
Ownership of radio	.435	1.076
Use of health center	.198	.615
Use of post office	.084	.187
Use of primary school	-.162	-.333
constant		-1.844

Table 17: Structure Matrix

Independent variables	Discriminant function loadings
Ownership of radio	.662
Education head of the household	-.587
Dependency ratio	.559
Source of cooking fuel	-.537
Toilet facility	.471
Source of water	-.413
Ownership of electric stove	.398
Use of post office	.387
Ownership of TV set	.332
Ownership of refrigerator	.262
Age of head of household	.230
Use of health center	.228
Ownership of sickle	-.209
Ownership of vehicle	.166
Ownership of bicycle	.161
Number of buildings owned	-.156
Ownership of farming animals	-.123
Ownership of farming animals	.123
Ownership of house	-.118
Gender of head of household	-.091
Use of primary school	.067
Ownership of land	-.060
Ownership of motorcycle	.040
Ownership of tractor	-.002

Table 18: Classification function coefficients

	Group Poor	Group Non Poor
Age head of the household	8.591	8.459
Ownership transport animals	2.942	3.259
Dependency ratio	1.542	1.180
Gender of the head of the household	10.111	10.411
Education head of the household	6.000	6.196
Number of buildings owned	9.557	9.741
Source of cooking fuel	4.428	4.567
Ownership of radio	19.494	18.746
Use of health center	7.947	7.520
Use of post office	9.616	9.486
Use of primary school	7.861	8.092
Constant	-82.528	-81.206

Table 19: Groups Means (Centroids) of canonical discriminant function

Group	Centroids
Poor	.406
Non Poor	-.289

Table 20: First classification matrix obtained by Proxy means test based on Discriminat Analysis

Observed	Predicted			% Correctly Classified
	Eligible	Ineligible	Total	
Poor	1025	462	1487	68.9%
Non poor	849	1190	2039	58.4%
Total	1874	1652	3526	62.8%

The probability estimated by the DA that a household belongs to the eligible group was used to test how the prediction could increase with different levels of probability selected a priory. It was decided to exclude households with a probability lower than .52 (2% higher than the chance) to belong to the eligible group. This lead to 12.3 percent of the households in the hold out sample being excluded, but a higher hit ratio for the remaining ones (Table 21).

Table 21: Second classification matrix obtained by Proxy means test based on DA

Observed	Predicted			% Correctly Classified
	Eligible	Ineligible	Total	
Poor	922	377	1299	71.0%
Non poor	706	1088	1794	60.6%
Total	1628	1465	3093	65.0%

The fact that 12 percent of the households had almost the same probabilities of being considered poor or non-poor confirms that undercoverage could be a problem if sensitivity of the test is low. Indeed, when the probability of belonging to a group was increased to at least 0.55, the proportion of the holdout sample excluded from the classification increased to 30.2 percent, but the hit ratio on the remaining sample increased to 66.8 percent.

These findings could help to reduce the two types of errors (undercoverage or leakage) by fixing different levels of probabilities to belong to one of the two groups. Households with a probability of being poor smaller than 0.4 or with a probability of not being poor smaller that 0.5 were excluded. This lead to 23.2 percent of the holdout sample being excluded but improve the hit ratio to 62.1 percent and the sensitivity to 85.3 percent. (Table 21)' with a drastic reduction of undercoverage with respect to the previous results (29% for Tab. 21; 31% for Tab. 20; 44.3% for Tab. 11; 45.1% for Tab 6).

Table 21: Third classification matrix obtained by Proxy means test based on Discriminant Analysis

Observed	Predicted			% Correctly Classified
	Eligible	Ineligible	Total	
Poor	1025	177	1202	85.3%
Non poor	849	657	1506	43.6%
Total	1874	834	2708	62.1%

3.3 Proxy Means Test by Logistic Regression

There are several reasons to consider logistic regression as an attractive alternative to DA. LR is less effected by heteroschedasticy (i.e., equalities of variances); it can be applied with categorical independent variables easily, whereas in discriminant analysis dummy variables can create problems; and it takes into account nonlinear relationships between the independent and the dependent variables, while this is not directly allowed by DA.

LR applies maximum likelihood estimation after transforming the dependent into a logit variable (the natural log of the odds of the dependent occurring or not). In this way, LR estimates the probability of a certain event occurring versus not occurring. Thus LR calculates changes in the log odds of the dependent, compared with the unit change in the dependent itself as in OLS. Independent variables can be nominal or ordinal and are often entered as dummy variables. SPSS can converts automatically categorical variables into dummies by leaving out the last or the first category against which the others are compared. Also in this case we used the same set of variables used in the previous analysis and the final model, obtained through stepwise procedure, is in tables 23 and 24.

The Wald statistics (Table 23) is commonly used to test the null hypothesis that a particular logit coefficient (effect) is zero. The significance level has the same meaning of the significance testing of b coefficients in OLS regression. Independents are usually dropped from the model (Table 24) when their effect is not significant by the Wald statistics.

Several measures are available to assess model fit. The -2LL (Table 25) decreases with the improvement of the model fit. The goodness of fit compares the predicted probabilities to the observed probabilities, with higher values indicating better fit. Of the two other measures, comparable with the multiple regression R^2, Nagelkerke's index is the most used because it varies between 0 and 1.

Table 23: Variables in the final LR model

Variable	B	SE	Wald	df	Sig	R	Exp (B)
Age of head of household (>65)			91.9121	4	.0000		.0837
Age 0 -| 15 (1)	-.9088	.7350	1.5287	1	.2163	.0000	.4030
Age 16 -| 25 (2)	.1948	.1244	2.4520	1	.1174	.0061	1.2151
Age 26 -| 35 (3)	-.2436	.0955	6.5027	1	.0108	-.0194	.7838
Age 36 -| 65 (4)	-.5250	.0855	37.7360	1	.0000	-.0546	.5916
Dependency ratio (>1.5)			283.880	3	.0000	.1523	
0 -| 0.5 (1)	1.0077	.0781	166.3813	1	.0000	.1171	2.7394
0.5 -| 1.0 (2)	.3436	.0780	19.3870	1	.0000	.0381	1.4100
1.0 -| 1.5 (3)	.0651	.0960	.4604	1	.4974	.0000	1.0673
Gender of head of household (male)	-.3325	.0619	28.8759	1	.0000	-.0474	.7171
Education head of the household (post secondary)			74.4021	5	.0000	.0733	
Illiterate (1)	.1573	.0734	4.5954	1	.0321	.0147	1.1704
Grade 1 - 3 (2)	.3723	.0945	15.5130	1	.0001	.0336	1.4510
Primary completed (3)	.5720	.1142	25.1061	1	.0000	.0439	1.7719
Junior secondary (4)	1.1968	.2144	31.1501	1	.0000	.0493	3.3094
Head secondary (5)	2.0211	.3801	28.2743	1	.0000	.0468	7.5467
Number of buildings owned (more than three)			13.8987	4	.0076	.0222	
none(1)	.3072	.3670	.7006	1	.4026	.0000	1.3596
one(2)	.4776	.3718	1.6501	1	.1989	.0000	1.6122
two(3)	.6954	.4058	2.9356	1	.0866	.0088	2.0044
three(4)	.8356	.4607	3.2900	1	.0697	.0104	2.3063
Source of water (own tap)			16.4977	5	.0056	.0233	
River/Lake (1)	.2403	.0661	13.2336	1	.0003	.0306	1.2717
Unprotected well (2)	.0429	.0739	.3364	1	.5619	.0000	1.0438
Other unprotected (3)	.0775	.1014	.5847	1	.4445	.0000	1.0806
Protected well (4)	.0703	.0875	.6447	1	.4220	.0000	1.0728
public tap (5)	-.1609	.1448	1.2342	1	.2666	.0000	.8514
Source of cooking fuel (electricity and other)			42.3143	6	.0000	.0503	
Leave/dung/other (1)	.2994	.0600	24.8914	1	.0000	.0437	1.3491
Collected wood (2)	.3699	.1122	10.8594	1	.0010	.0272	1.4475
Purchased wood (3)	.7898	.2474	10.1860	1	.0014	.0261	2.2028
Charcoal (4)	.7575	.1659	20.8512	1	.0000	.0397	2.1330
Kerosene (5)	1.1735	.5758	4.1540	1	.0415	.0134	3.2334
Gas (6)	.3089	.2649	1.3603	1	.2435	.0000	1.3619
Toilet facility (other)			11.3417	4	.0230	.0167	
Flush Toilet (1)	-.4544	.2932	2.4019	1	.1212	-.0058	.6348
Pit (2)	-.4053	.1850	4.8009	1	.0284	-.0153	.6668
Container (3)	-.7241	.6521	1.2328	1	.2669	.0000	.4848
Field/Forest (4)	-.5236	.1675	9.7645	1	.0018	-.0255	.5924
Ownership of radio (yes)	.8047	.0796	102.125	1	.0000	.0914	2.2359
Ownership of TV set (yes)	.5777	.2156	7.1768	1	.0074	.0208	1.7819
Ownership of sickle (yes)	.1167	.0586	3.9652	1	.0464	.0128	1.1237
Use of health center (yes)	.4243	.0718	34.9132	1	.0000	.0524	1.5285
Use of post office (yes)	-1.1010	1.1370	.9376	1	.3329	.0000	.3325
Use of primary school (yes)	-.2132	.0537	15.7887	1	.0001	-.0339	.8080
Constant	.6544	1.2119	.2915	1	.5892		

Table 24: Variables not in the equation of the LR analysis

Variable	Score	df	Sig	R
Ownership of house	5.5532	3	.1355	.0000
Ownership of land	.5370	1	.4637	.0000
Ownership of vehicle	.4462	1	.5041	.0000
Ownership of refrigerator	2.5095	1	.1132	.0065
Ownership of electric stove	.1878	1	.6647	.0000
Ownership of tractor	.1782	1	.6729	.0000
Ownership of farming animals	.0346	1	.8525	.0000
Ownership of bicycle	1.3230	1	.2501	.0000
Ownership of motorcycle	.5998	1	.4387	.0000

Table 25: fit of the LR model

-2 Log Likelihood	10769.357
Goodness of Fit	8761.213
Cox & Snell - R^2	.128
Nagelkerke - R^2	.173

The testing of the model (Table 26) on the holdout sample produced a hit ratio of 66.1 percent and was more efficient in correctly identifying the poor (76%) than the non poor (52.4%).

Table 26: First classification matrix obtained by proxy means-test based on LR analysis

Observed	Predicted			% Correctly Classified
	Eligible	Ineligible	Total	
Poor	779	708	1487	52.4%
Non poor	486	1553	2039	76.2%
Total	1265	2261	3526	66.1%

As for the DA, the LR allowed to compute the probability of a household belonging, alternatively, to the poor or non-poor groups. Thus, also LR could be applied

to improve flexibility and efficiency of targeting. Applying a probability equal or greater than .53 (3 percent greater than chance) produced an exclusion of 10 percent of households and a higher hit ratio in the rest. (Table 27).

Table 27 11[vi]: Second classification matrix obtained by Proxy means test based on LR

Observed	Predicted			% Correctly Classified
	Eligible	Ineligible	Total	
Poor	670	612	1282	52.3%
Non poor	394	1421	1815	78.3%
Total	1064	2033	3097	67.5%

A probability of being poor lower than 0.4 or the probability of being non-poor lower than 0.35 was accompanied to a 50 percent of the hold out sample left out. The rest of the sample had a hit ratio of 73 percent, a proportion of poor correctly identified of 81 percent and a proportion of poor correctly classified of 60 percent (Table 28).

Table 28: Third classification matrix obtained by proxy means-test based on LR

Observed	Predicted			% Correctly Classified
	Eligible	Ineligible	Total	
Poor	402	267	669	60.1%
Non poor	209	885	1094	80.9%
Total	611	1152	1763	73.0%

3.4 An alternative strategy to define poverty

An alternative way to the money metric and in line with the philosophy of the proxy means-test is to detect households with similar welfare characteristics through:
(*i*) The Non-Linear Principal Component Analysis (NLPCA), which transforms the original indicators into optimally quantified and standardized composite dimensions; and
(*ii*) The Cluster Analysis (CA), which is applied on the NLPCA dimensions to partition the households into clusters with a high homogeneity within each cluster and a high heterogeneity among clusters.

3.4.1 NLPCA

The NLPCA produced new variables or dimensions that summarized the information scattered across many variables. The first six dimensions explained 59 percent of the total variance and the characteristics of each dimension depended on the Factor Loadings (Table 29). These are measures of correlation of the original variables with the dimensions and therefore help to give a meaning to the dimensions. The first dimension was characterized by several proxies of welfare such as: education of the head of the household; ownership of land, farming and transport animals, house, radio, electric stove, sickle, tractor; source of water and cooking fuel; sanitation facilities; and access to services such as the post office. . It was a measure derived from the combination of the original variables, with negative values for the poorest households ([9]), and positive values for the better off households ([10]). The second dimension was characterized by possession of "luxury items" such as TV, refrigerator and vehicles. Gender and education of the head of the households and ownership of transport and farming animals characterized the third dimension. The other dimensions were progressively less important because they explained a lower proportion of the total variance.

([9]) For example, -1.28 corresponded to a household with: illiterate head and 45 years old; five children (2 male and 3 female); total income and expenditure of 993 and 289 birrs; no radio, vehicle, tractor, TV; ownership of farming animals and sickle; use of unprotected water and collected wood.
([10]) For example, 2.80 corresponds to a 35 years old head of the household with post sec. education, two children (1 male and 1 female), total income and expenditure of 11472 and 2959 birrs; with radio, motorcycle, TV, electric stove; use of protected water and electricity.

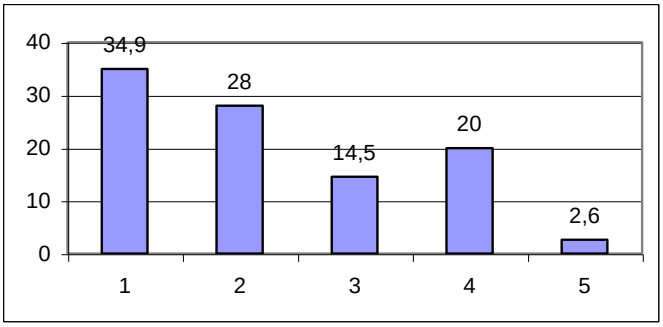
Table 29: Factor loadings of the NLPCA.

Variables	Dimension					
	1	2	3	4	5	6
Age of the head of the household	.039	.075	.319	**.654**	-.242	.248
Ownership of transport animals	**.657**	.332	**-.428**	.021	-.377	.156
Dependency ratio	197	.034	-.104	**.488**	-.224	-.183
Gender of the head of the household	-.252	-.368	**.462**	.165	-.167	-.017
Education head of the household	**-.594**	.230	**-.411**	-.253	.221	-.041
Number of buildings owned	**-.297**	-.167	-.069	-.074	-.207	-.006
Source of water	**-.815**	.039	-.033	.108	-.003	.103
Source of cooking fuel	**-.851**	.012	-.075	.017	-.057	.108
Toilet facilities	**.793**	-.074	.116	-.082	-.019	-.108
Ownership of house	**-.682**	-.378	-.161	-.183	**-.456**	-.018
Ownership of land	**-.658**	-.385	-.148	-.177	**-.470**	-.013
Ownership of a vehicle	.225	**-.559**	-.305	.124	.174	.112
Ownership of radio	**.683**	-.245	.219	-.051	-.080	-.086
Ownership of TV set	**.414**	**-.607**	-.250	.060	.025	-.180
Ownership of refrigerator	.341	**-.601**	-.293	.071	.076	-.165
Ownership of electric stove	**.581**	-.304	-.015	-.009	.010	-.143
Ownership of sickle	**-.650**	-.314	.179	-.039	-.009	-.027
Ownership of tractor	**.681**	.105	.052	-.174	-.027	.075
Ownership of farming animals	**-.640**	-.329	**.432**	-.014	.392	-.164
Ownership of bicycle	.176	-.368	-.082	.095	.040	**.453**
Ownership of motorcycle	.037	-.279	-.181	.124	.336	**.612**
Use of the health center	.122	-.030	.295	-.309	-.212	**.425**
Use of post office	**.574**	-.145	.265	-.145	-.128	-.011
Use of the primary school	.361	-.112	.204	**-.535**	-.028	.164

Loadings larger than |0.4| are evidenced and considered sufficient contributions to identify dimensions

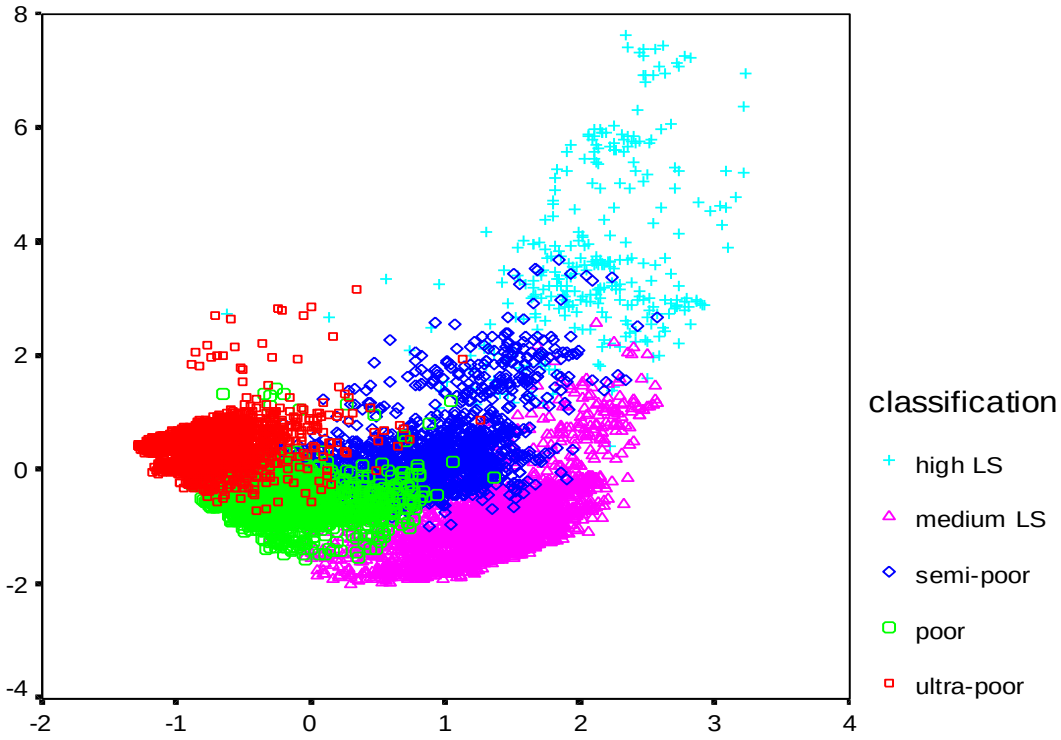
The CA used the NLPCA dimensions to form five homogeneous clusters through the *k*-means algorithms. These clusters were ranked according to the average score of the first dimension of NLPCA. The frequency

Figure 6: frequencies of the households in the five clusters



The location of the clusters in the space defined by the first two NLPCA dimensions is shown in Figure 7. Because higher values of the dimensions were associated with better standards of living, the better off households were located on the upper right hand side and worst off households on the bottom left hand side.

Figure 7: location of the households in the plane spanned by the first two dimensions of the NLPCA.



3.5 Comparison between the expenditure quintiles and the clusters

The clusters and the expenditure quintiles were compared in terms of prevalence of indicators of deprivation. Because of the different metrics used, the classification of poor households according to the CA is not directly comparable with the expenditure quintiles. However, because both classifications have the same objective of identifying deprivation, they were validated against welfare variables and biological deprivation measured through stunting and wasting among children less than 5 years old.

Stunting is the most holistic biological index of the deprivation. It is the result of slow linear growth, resulting in a failure to achieve expected height compared to a child of the same age living in a healthy environment. It is also known as chronic malnutrition because it is a long lasting effect of long term deprivation from chronic insufficient protein energy intake, frequent infections, sustained incorrect feeding practices and low socioeconomic family status.

Wasting occurs when a child's weight falls significantly below what is expected of a child of the same height living in a healthy environment. Wasting indicates current acute malnutrition resulting from failure to gain weight or actual weight loss. Causes include inadequate food intake, incorrect feeding practices, infections and/or a combination of these factors. Wasting in individual children and population groups can change rapidly and shows marked seasonal patterns associated with changes in food availability or disease prevalence to which it is very sensitive (Beaton et al (1990)).

Compared with expenditure quintiles, the clusters were characterized by a higher variation in illiteracy, use of unprotected wells, type of cooking fuel, ownership of items and dependency ratio (Figures 8-17). This higher characterization of the cluster was also evident for the biological deprivation; with stunting prevalence improving from 44 to

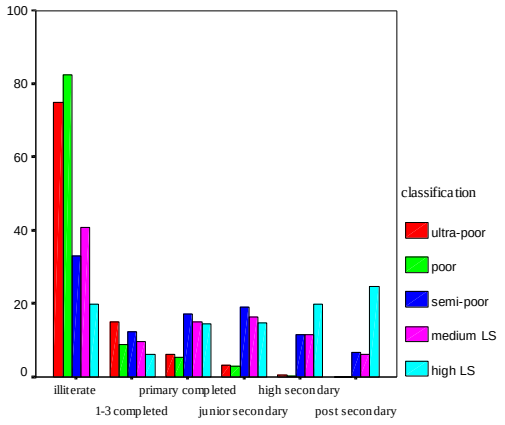
20 percent between the first and fifth cluster compared with 47 to 32 percent between the first and fifth expenditure quintiles (Table 30). The situation was even more striking with wasting which declined from 10 to 4 percent between the first and fifth quintiles; while there was no clear pattern for the expenditure quintiles (Table 31). These findings suggest that in countries like Ethiopia, NLPCA and CA can provide a powerful tool to differentiate groups characterized by different standard of living.

Table 30 Average per capita total household expenditure by cluster

Cluster 1 Ultra –poor	Cluster 2 Poor	Cluster 3 Semi-poor	Cluster 4 Medium LS	Cluster 5 High LS
1152	1222	1558	1557	2369

Figure 8: Education of the head by: a) clusters and b) expenditure quintiles

a) Classes of households



b) Quintiles of Expenditure

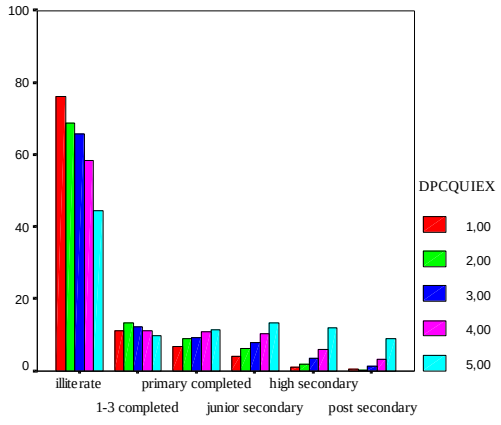


Figure 9: Type of water used by: a) clusters and b) expenditure quintiles

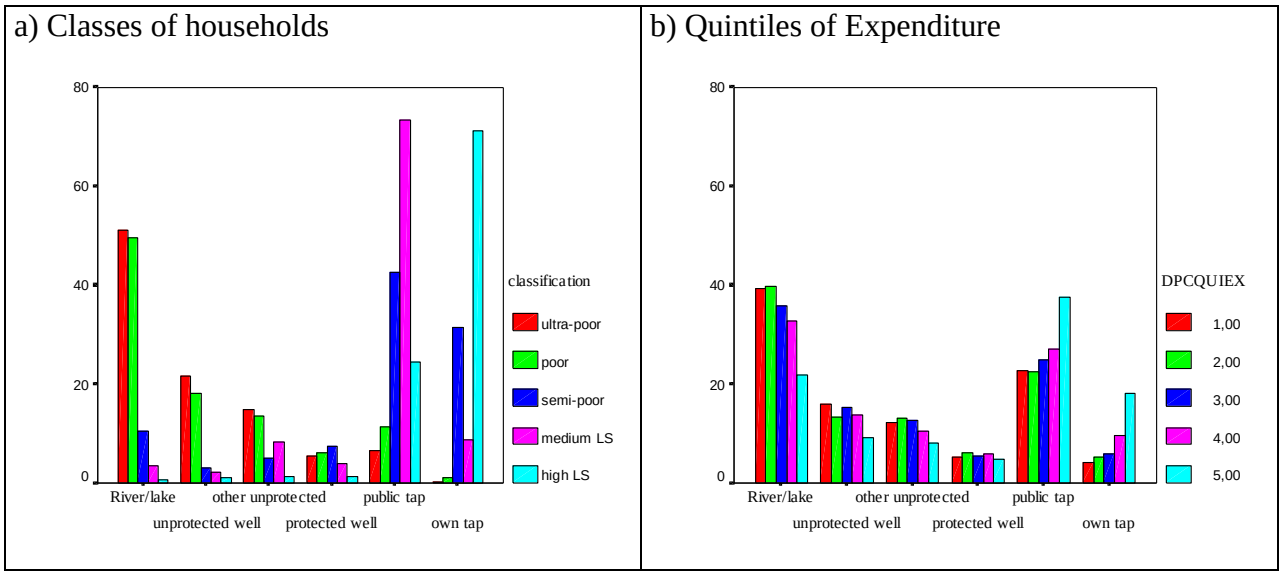


Figure 10: Source of cooking fuel by: a) clusters and b) expenditure quintiles

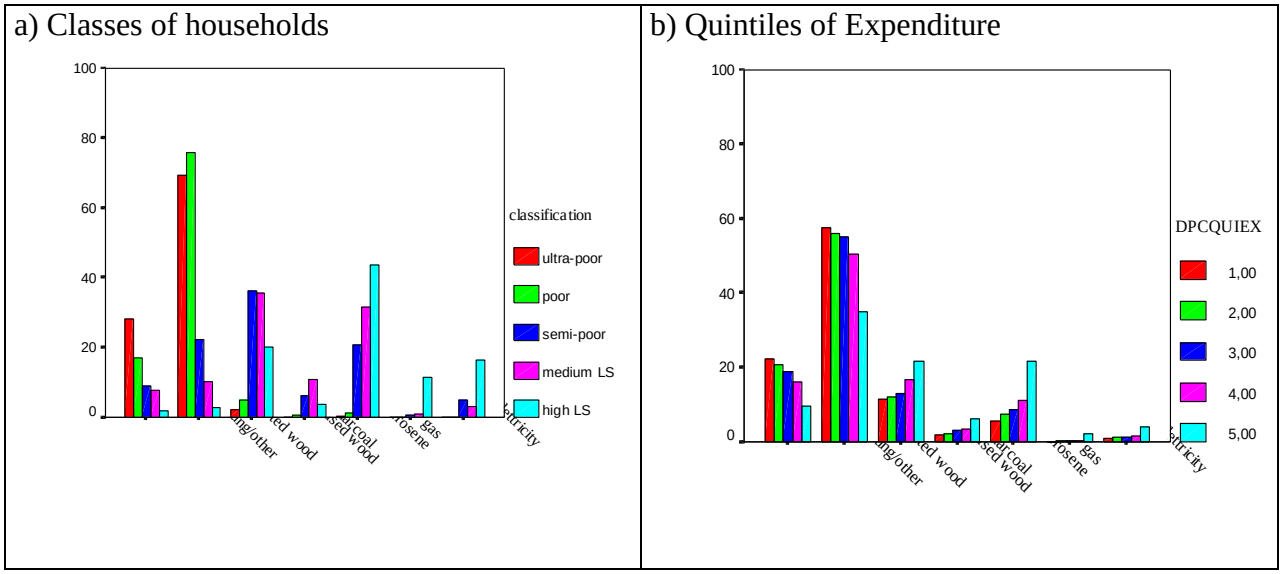


Figure 11: Ownership of land by: a) clusters and b) expenditure quintiles

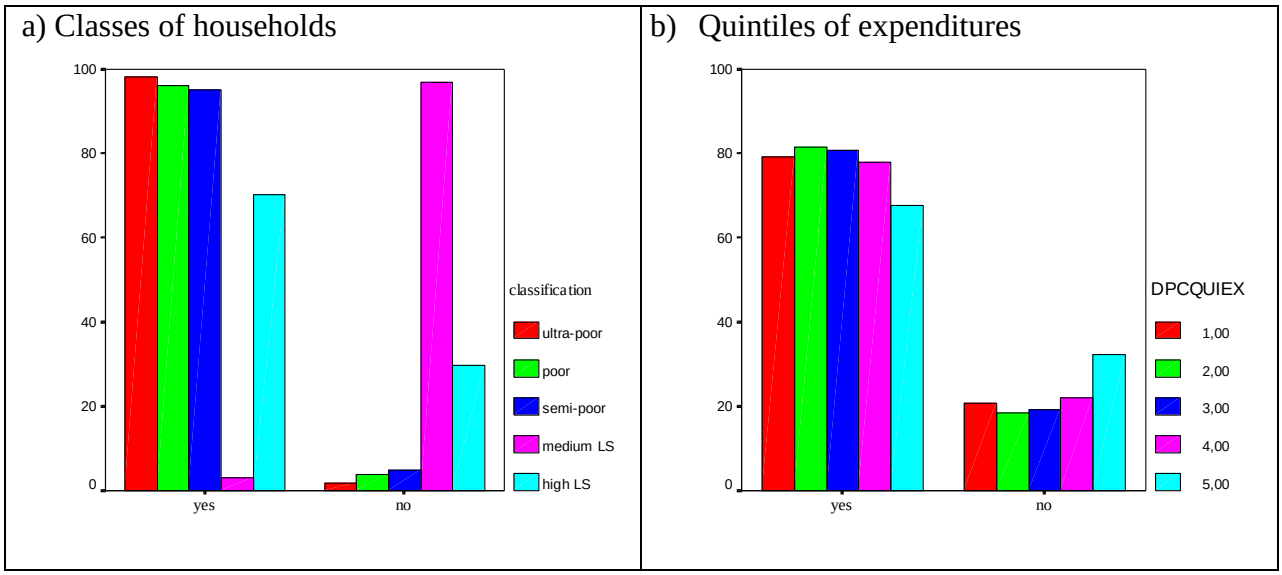


Figure 12: Ownership of TV by: a) clusters and b) expenditure quintiles

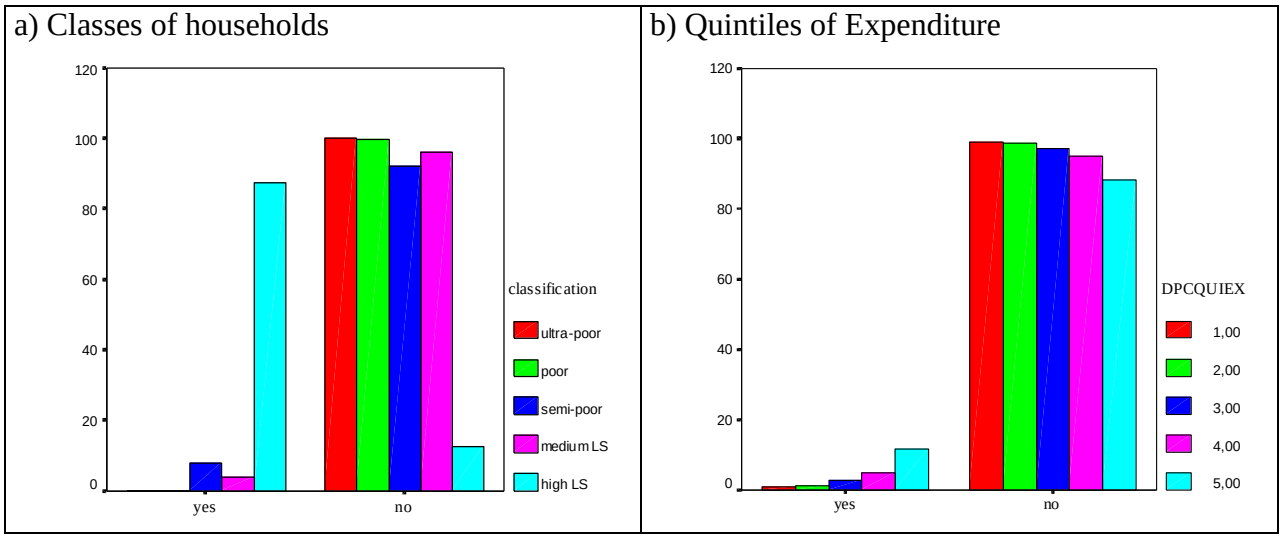


Figure 13: Ownership of vehicle by: a) clusters and b) expenditure quintiles

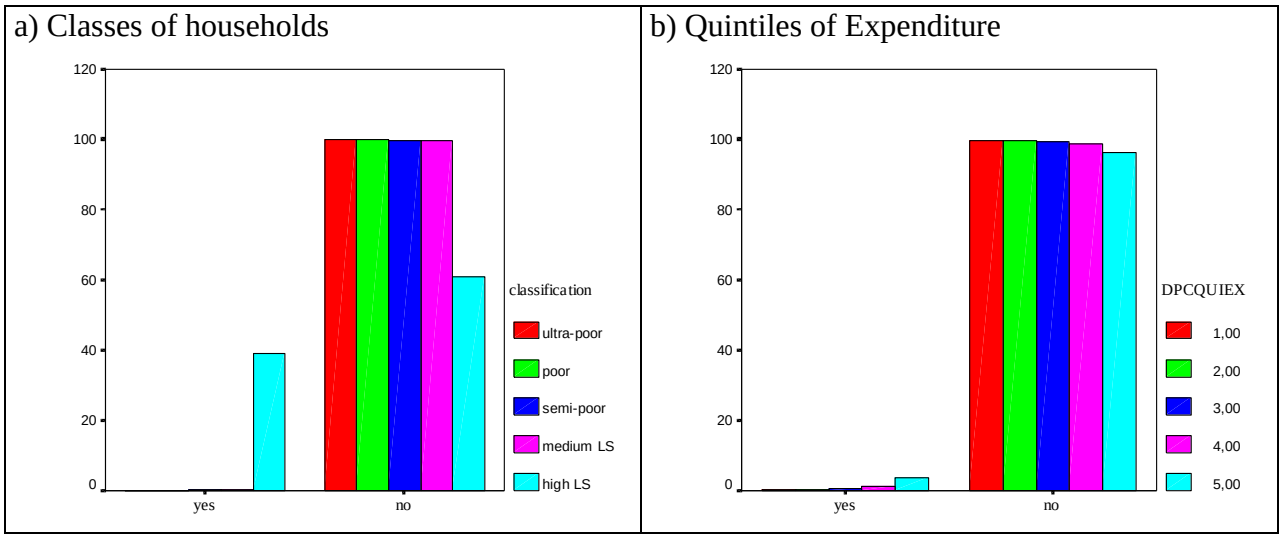


Figure 14: Ownership of a fridge by: a) clusters and b) expenditure quintiles

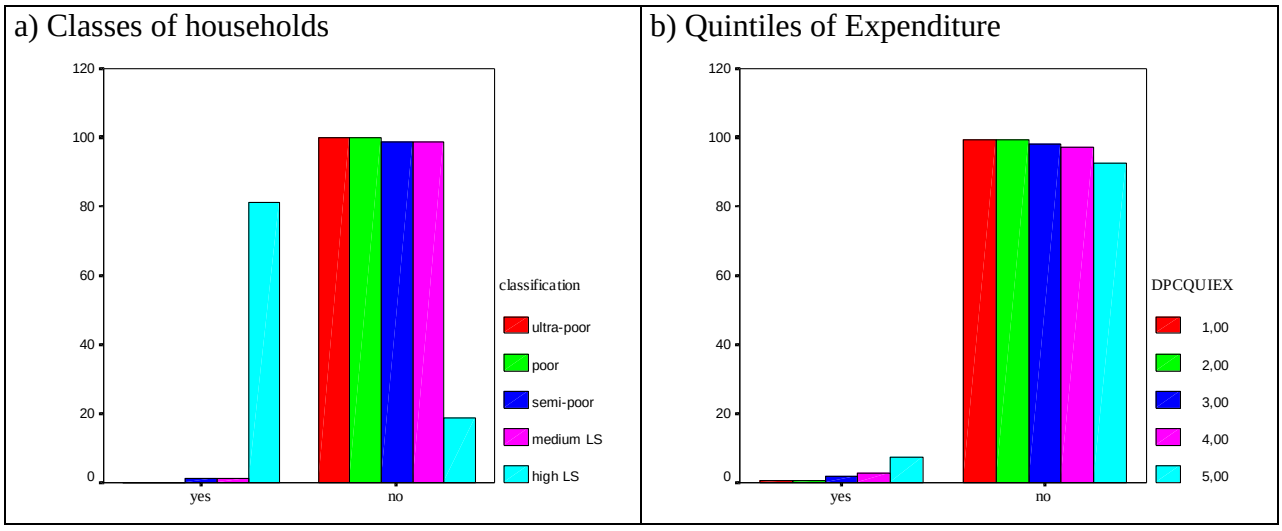


Figure 15: Ownership of radio by: a) clusters and b) expenditure quintiles

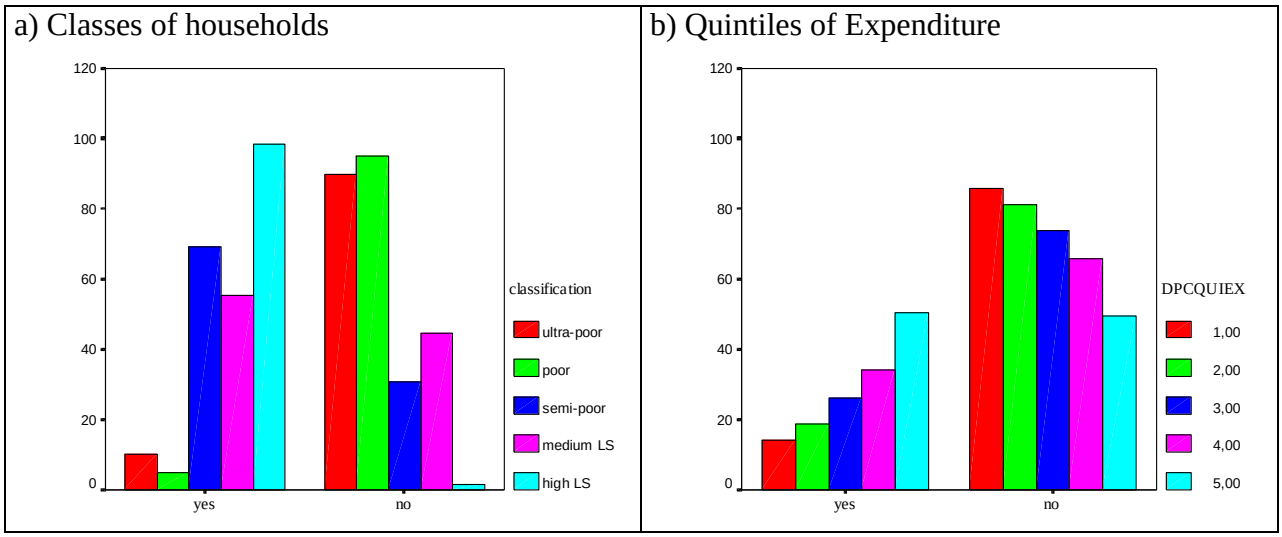


Figure 16: Ownership of stove by: a) clusters and b) expenditure quintiles

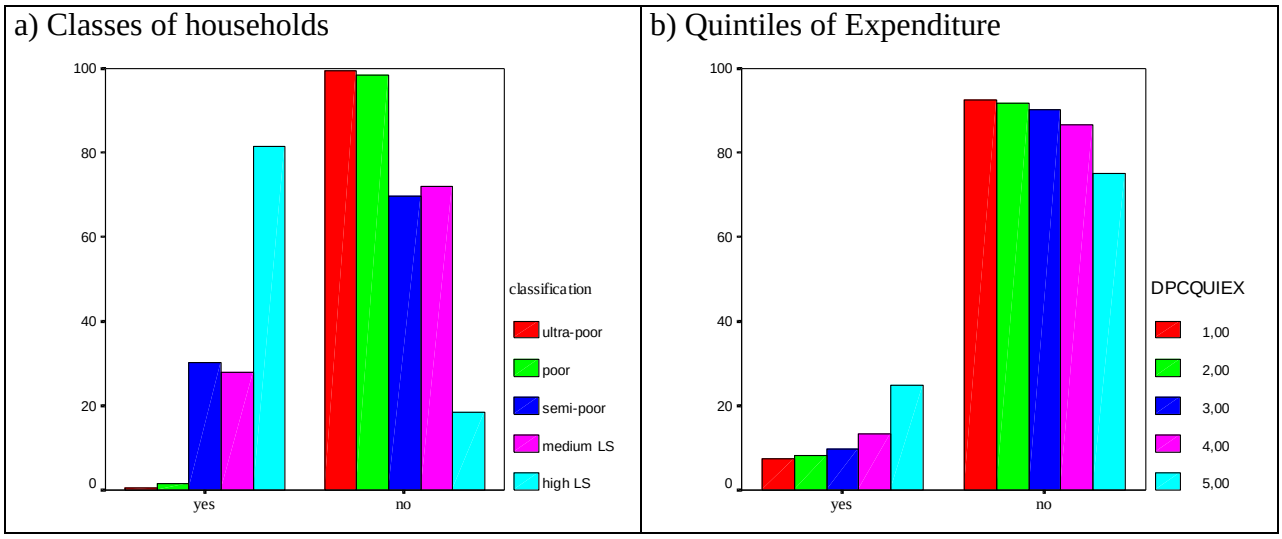


Figure 17: dependency ratio by: a) clusters and b) expenditure quintiles

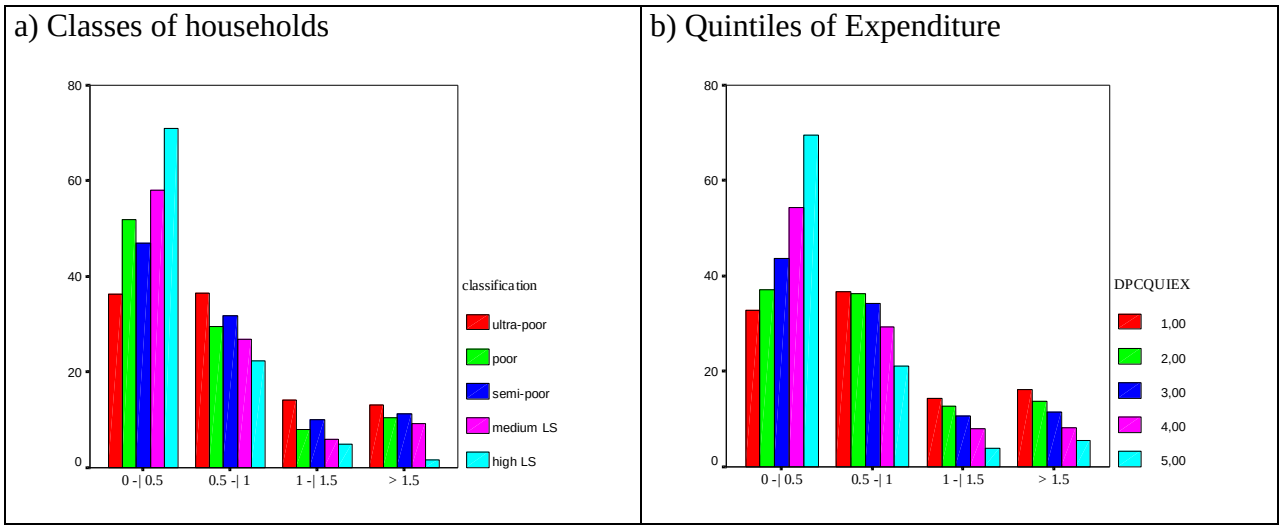


Table 30: *Prevalence of seriously stunted children under five years of age (*).*

(*a*)	Ultra-Poor	Poor	Semi-Poor	Medium LS	High LS
CA	44.3	44.9	37.1	31.8	20.2

(*b*)	I quintile	II quintile	III quintile	IV quintile	V quintile
Quintiles of expenditure	46.9	45.8	42.7	39.1	31.8

(*) *Whose height for age was below -3 standard deviations of the reference standard (NCHS)*

Table 31: *Prevalence of wasted children under five years of age (*).*

(*a*)	Ultra-Poor	Poor	Semi-Poor	Medium LS	High LS
CA	9.9	8.9	5.7	6.0	4.1

(*b*)	I quintile	II quintile	III quintile	IV quintile	V quintile
Quintiles of expenditure	8.1	10.7	8.7	8.0	9.0

(*) *Whose weight for length was below -2 standard deviations of the reference standard (NCHS)*

4. Conclusions

The predictive accuracy of the proxy means test is measured by the hit ratio, which is the proportion of households correctly classified. The researcher may ask what can be considered a desirable level of predictive accuracy for the targeting mechanism, with respect to the percentage of correct classifications obtained by chance (no targeting). If poor households are defined according to the 40^{th} percentile of expenditure and no targeting mechanism is applied a 40 percent 'classification accuracy' is achieved (*minimum chance criterion*). If, as absurd as it may seem, it is decided that all the households are ineligible, the accuracy of the classification rises to 60 percent (*maximum chance criterion*). However, a means-test mechanism is used to correctly classify both groups (poor and non-poor), and therefore another chance model has to be used which is called *proportional chance.* The formula for this criterion (Huberty, C.J., Winsenbaker, J.W. & Smith J.C., 1987) is Cpro= p2+ $(1-p)^2$, where *p* is the proportion of the poor. If *p is* 0.4, the proportional chance accuracy is 52 percent, which is corresponding to the hit ratio.

The hit ratio and the proportion of poor correctly identified were respectively 62.3 percent and 55 percent by the proxy means test based on OLS regression. When outliers[11] were deleted, these proportions rose to 62.6 and 56 percent. Exclusion of outliers helped to reduce serious violations of the regression assumptions, but did not strongly change the predictive power of the hit ratio. This first result, worthy further investigation, shows the substantial robustness of OLS in testing eligibility or ineligibility of households, even under moderate violation of the regression assumptions.

The accuracy improved slightly with the DA, which produced a hit ratio and a proportion of poor correctly identified of respectively 62.8 and 69 percent. The probabilities produced by the DA were tested to show how targeting may be used to reduce undercoverage and/or leakage.

[11] observations outside +3 studentized residuals

LR seemed more efficient in terms of hit ratio (66.1%) and in correctly identifying the non poor (76.2%) but it was less efficient in correctly identifying the poor (52.4%). Also LR offered the possibility to compute probabilities to belong to the poor and non-poor groups to reduce leakage and/or undercoverage.

All the analytical methods used in the study have shown a significantly larger percentage of corrected classifications than would be expected by proportional chance model (52%), that is, using random targeting. The novelty introduced by the DA and LR was the possibility of assigning different probabilities to classify households to the poor and non poor group. These encouraging results justify further studies to test the applicability of these analytical techniques to improve the accuracy and flexibility of proxy means-test in reducing undercoverage and/or leakage.

Finally, the positive findings of the NLPCA and CA confirmed their potential in providing a more holistic approach to define poverty. This is not a minor advantage, considering that the data available from expenditure surveys are much scarcer than the welfare indicators collected in most surveys. This justifies an expanded use of alternative analytical methods to capture the essence of poverty in its multidimensional aspects. The promising results of the NLPCA and CA in capturing such aspects should be taken into consideration by the 2000 World Development Report of the World Bank.

References

Barnett, V., & Lewis T. (1984), *Oultiers in Statistical Data*, 2nd edition, New York: Wiley.

Beaton, G., Kelly, A., Kevany, J., Martorell, R., & Mason, J., (1990), *Appropriate Uses of Anthropometric Indices in Children*. Geneva: UN ACC/SCN.

Dillon, W. R., & Goldstein M. (1984), *Multivariate Analysis: Methods and Applications*, New York, Wiley.

Datt, G., and Ravallion M. (1993), "Regional Disparities, Targeting and Poverty in India." In Lipton M and Van der Gaag J. eds. Including the poor.

Everitt, B., (1993), *Cluster Analysis*, 3rd edition, New York, Halsted Press.

Gordon, A. D. (1987). A Review of Hierarchical Classification, The Journal of the Royal Statistical Society, A, 150, 2, 119-137.

Green, P. E., and J.D. Carroll (1978), Mathematical Tools for Applied Multivariate Analysis, New York: Academic Press.

Grosh, M. (1994), Administering Targeted Social Programs in Latin America: from Platitudes to Practice, World Bank.

Grosh, M. & Backer, J., (1995), Proxy Means Tests for Targeting Social Programs: Simulations and Speculations, Living Standard Measurement Study, Working Paper 118, World Bank.

Grosh, M. & Glinskaya, E., (1998), Proxy Means Testing and Social Assistance in Armenia, Draft for discussion, World Bank.

Haddad L. et al., (1991), Identification and evaluation of alternative Indicators of Food and Nutrition Security: Some Conceptual Issues and an analysis of Extent Data." IFPRI, Washington DC.

Hosmer, D. W., & Lemeshow S. (1989), *Applied Logistic Regression*, New York, Wiley.

Hubert, C. J., Wisenbaker, J. W., & Smith J.C. (1987), Assessing Predictive Accuracy in Discriminant Analysis, Multivariate Behavioral Research 22, 307-329.

Kaufman, L. and J. Rousseeuw P., (1990), *Finding groups in data: An introduction to cluster analysis* NY: John Wiley & Sons.

Ravallion M., (1993), "Poverty Alleviation Trough Regional Targeting: A Case Study for Indonesia." In Hoff K. et al The Economics of Rural Organization: Theory, Practice and Policy. The World Bank.

Ravallion, M. & Chao, K., (1989), Targeted Policies for Poverty Alleviation Under Imperfect Information: Algorithms and Applications, The Journal of Policy Modeling, 11, 2, 213-224.

Singer, P. (1983), *Practical Ethics*, Cambridge University Press.

United Nations Development Programme (1997), *Human Development Report*, Oxford University Press.

World Bank (1997), *World Development Indicators*, Washington, DC

Methodological Annex

1 *Discriminant Analysis*

Overview

The DA used welfare proxies, from poor and non poor households, to build a weighted function that could predict eligibility in the context of means testing. DA assumes that the dependent variable is truly dichotomy and shares all the usual assumptions of correlation. These include linear and homoscedastic relationships, and untruncated interval or near interval data. Like multiple regression, DA proper model specification with inclusion of all the relevant independents and exclusion of redundant variables to improve the fit of the model.

1.1 *Model selection*

To apply DA it is necessary to specify a binary dependent variable that identifies two mutually exclusive groups. In this study the DA was applied to the analysis sample which was categorized into poor and non-poor households, according to the 40^{th} percentile of expenditures. The independent variables were those already identified through the OLS.

Two computation methods can be utilized in selecting a Discriminant function: the simultaneous and stepwise method. The first one involves computing the discriminant function with all the independent variables in the model. The simultaneous method is appropriate when the researcher has to verify a theoretical model. Stepwise involves entering the independent variables into the model on the basis of their discriminating power. The stepwise method select sequentially the best discriminating variable at each step, eliminating variables that are not useful in discriminating between the two groups defined by the dependent variable.

1.2 *Discriminant function*

DA creates deriving a new variable (c*riterion variable)*, that is a linear combination of the independent variables (*predictors)*, to discriminate best between the two dichotomous groups. Discrimination is achieved by setting the weights of the linear combination for each variable to maximize the between group variance relative to the within-group variance. The linear combination for a DA, called discriminant function is equal to:

$$z_{ik} = a + w_1x_{1k} + w_2x_{2k} + \ldots + w_nx_{nk}$$

where

z_{ik} is the discriminant z score of discriminant function j for object k;

a is the intercept base line;

w_i is the discriminant weight for the independent variable i;

x_{ik} is the independent variable i for object k.

The discriminant function is analogous to the multiple regression function, but the w's are discriminant coefficients that maximize the distance between the means of the criterion (dependent) variable. Although the discriminant function is estimated using ordinary least-squares, maximum likelihood estimation is also used.

The *discriminant score*, is the value resulting from applying a discriminant function formula to the data of a given case. The *Z score* is the discriminant score for standardized data.

The test for the statistical significance of the discriminant function is a generalized measure of the distance between the group centroids, which correspond to the mean of the discriminant z score for all the objects within each group. The tests to evaluate the statistical significance of the discriminant function include Wilks' lambda, Hotelling's trace, and Pillai’s criterion.

If a stepwise method is used to estimate the discriminant function, the Mahalanobis D^2 and Rao's V measures are the most appropriate. The Mahalanobis D^2 may become critical as the number of predictors increase. The conventional significant level of 0.05 is often used.

Wilks's lambda is used in an *ANOVA (F) test of mean differences* in DA, such that the smaller the lambda for an independent variable, the higher the variable contribution to the discriminant function. Lambda varies from 0 to 1, with 0 meaning group means differ (thus the more the variable differentiates the groups), and 1 meaning all group means are the same. The F test of Wilks's lambda, also known as *the U statistic*, is used to test the significance of variables' contributions and of the discriminant function as a whole.

Once the significant discriminant function has been identified, it is necessary to ascertain the overall fit of the model. This may be obtained using a test such as Mahalanobis D^2, which however does not tell how well the function predicts. For this reason, it is customary to develop classification matrices to provide a more accurate assessment of the discriminating power of the function. A classification matrix is a crosstabulation of actual group membership with predicted group membership. The frequencies in the main diagonal represent correct classifications, while the off-diagonal frequencies represent incorrect classifications. The percentage of cases correctly classified is called the *hit ratio*.

To classify it is necessary to determine the cutting scoring (cut-off), against which each object is classified. If the discriminant score of the function is less than or equal to the cut-off, the case is classed as 0, otherwise it is classed as 1. When group sizes are equal, the cutoff is the mean of the two centroids (for two-group DA); otherwise it is the weighted mean.

1.3 *Interpreting the results of DA*

The *structure matrix table* shows the correlation of each variable with each discriminant function. These simple Pearsonian correlations also known as *structure*

coefficients, correlations or *discriminant loadings. The structure coefficients* measure the simple correlations between the variables and the discriminant function. Therefore the structure coefficients identify which variables characterize most the discriminant function.

The standardized discriminant function coefficients indicate the partial contribution of each variable to the discriminant function controlling for other independent variables in the equation. These are used to compare the relative importance of the independent variables, much as beta weights are used in regression. The standardized discriminant function coefficients should be used to assess each independent variable's unique contribution to the discriminant function.

Unstandardized discriminant coefficients are used in the formula for making the classifications in DA, much as b coefficients are used in regression in making predictions. The discriminant scores are obtained through the product of the unstandardized coefficients with the observations' characteristics.

The *group centroid* is the mean value for the discriminant scores for a given category of the dependent, which being binomial has two centroids, one for each group.

At the beginning of the analysis the sample is divided into the analysis sample to build the function and the *hold-out sample* which is used to validate the discriminant function by assessing its performance on correctly classifying each case of the hold-out sample.

Mahalanobis distances are used to analyze a new, unknown set of cases in comparison to an existing set of known cases. Mahalanobis is the distance between a case and the centroid for each group (of the dependent) in attribute space (n-dimensional space defined by n variables). A case will have one Mahalanobis distance for each group, and it will be classified as belonging to the group for which its Mahalanobis distance is smallest. Thus, the smaller the Mahalanobis distance, the closer the case is to the group centroid and the more likely it is to be classified as belonging to that group. Since

Mahalanobis distance is measured in terms of standard deviations (SD) from the centroid, a case which is less than 1.96 Mahalanobis distance units from the centroid has less than .05 chance of belonging to the group represented by the centroid. Three SD units would likewise correspond to less than .01 chance. SPSS reports squared Mahalanobis distance.

Tests of Assumptions. Box's M tests the null hypothesis that the covariance matrices do not differ between groups formed by the dependent. If this test is not significant, the null hypothesis that the groups do not differ is accepted. Although the test is very sensitive in meeting also the assumption of multivariate normality, DA can be robust even when this assumption is not met.

1.4 *Assumptions of DA*

Homogeneity of variances (homoscedasticity): within each group formed by the dependent, the variance of each interval independent should be similar between groups. That is, the independents may (and will) have different variances one from another, but for the same independent, the groups formed by the dependent should have similar variances and means on that independent.

Each group has a similar covariance/correlation matrix. Homogeneity of covariances/correlations: within each group formed by the dependent, the covariance/correlation between any two-predictor variables should be similar to the corresponding covariance/correlation in other groups.

Low multicollinearity of the independents. To the extent that independents are correlated, the standardized discriminant function coefficients will not reliably assess the relative importance of the predictor variables.

Linearity and additivity. DA does not take into account exponential terms unless such transformed variables are added as additional independents. DA does not take into account interaction terms unless new crossproduct variables are added as additional independents).

For purposes of significance testing, predictor variables follow multivariate normal distributions, with each predictor having a normal distribution about fixed values of all the other independents.

2. *Logistic regression*

Overview

IN LR the dependent variable is dichotomous; and the independents are continuous and categorical variables. LR may be preferred with respect to DA for several reasons including:

- DA relies often on meeting the assumptions of multivariate normality and equal variance-covariance matrices across groups - assumptions that may not be met in some situations. LR does not face these strict assumptions and it is much more robust when these assumptions are not met.
- LR takes into account nonlinear relationships between the dependent and the predictive variables.
- LR can handle categorical independent variables easily, whereas in DA the use of dummy variables creates problems with the variance/covariance equalities.
- The interpretation of the results and of the diagnostic measures for examining residuals is similar to those of multiple regression.

LR approaches the prediction of group membership directly determining the probability of an object to belong to one of the two groups. To define a probability between 0 and 1, LR uses an assumed relationship between the independent and the dependent variables that resemble an S-shaped curve. At the very low levels of the independent variables, the probability (dependent variable) is near zero. As the independent variables increases the probability increases, up to a point when the slope starts decreasing to approach one, but never exceeding it.

2.1 *Estimating Logistic Model*

Multiple regression employs ordinary least squares approach to minimize the sum of squared differences between the actual and the predicted values of the dependent

variable. The nonlinear nature of the logistic function requires that the maximum likelihood procedure, be used in an iterative manner to find the most likely estimates for the coefficients. This results in the use of the likelihood value instead of the sum of squares when calculating measure of overall model fit.

Logit coefficients, also called *effect coefficients*, correspond to the *b* (unstandardized regression) coefficients in OLS regression, and are used in the LR equation to estimate (predict) the values of the dependent. If the logit for a given independent variable is b_1, then a unit increase in the independent variable is associated with b_1 unit increase in the log odds of the dependent variable (the natural log of the probability that the dependent = 1 divided by the probability that the dependent = 0). Therefore, LR calculates changes in the log odds of the dependent, and it does not estimate changes in the dependent itself as OLS regression does.

LR derives its name from the logistic transformation. The logistics curve of the LR predicts that an object is likely to belong to the first group if the probability is greater than 0.5. The logit can be converted easily into odds *ratio* of the dependent rather than log odds by using the exponential function (raising the natural log to the b_1 power). For instance, if the logit $b_1 = 2.303$, then its log odds ratio (exponential function) is 10, indicating that for one unit increase of the independent variable the odds that the dependent = 1 increases by a factor of 10, when other variables are controlled for.

If the Beta is positive, its transformation (antilog) will be greater than one, and the odds will increase. This increase occurs when the predicted probability of the event's occurring increases and the predicted probability of its not occurring is reduced. Likewise, if Beta is negative, the antilog is less than one and the odds will decrease. A coefficient equal zero is equal to one, resulting in no change in the odds.

Maximum Likelihood Estimation (MLE) is the method used to calculate the logit coefficients. This differs from the OLS estimation of coefficients by minimizing the sum of squared distances of the data points to the regression line. MLE seeks to maximize the log likelihood (LL), which reflects how likely it is (the odds) that the observed values of

the dependent, coded as 0 or 1, may be predicted from the observed values of the independents.

MLE is an iterative algorithm that starts with an initial arbitrary logit; the MLE algorithm determines the direction and size change in the logit coefficients that will increase LL. After this initial function is estimated, the residuals are tested and a re-estimate is made with an improved function, and the process is repeated (usually about a half-dozen times) until *convergence* is reached (that is, until LL does not change significantly).

Log likelihood. A "likelihood" is a probability, specifically the probability that the observed values of the dependent may be predicted from the observed values of the independents. Like any probability, the likelihood varies from 0 to 1. The log likelihood (LL) is its log and varies from 0 to minus infinity (it is negative because the log of any number less than 1 is negative). LL is calculated through *iteration*, using a maximum likelihood method. Because -2LL has approximately a chi-square distribution, -2LL can be used to assess the significance of LR, analogous to the use of the sum of squared errors in OLS regression.

R-squared. There is no widely accepted direct analog to OLS regression's R^2. This is because an R^2 measure seeks to make a statement about the "percent of variance explained," but the variance of a dichotomous or categorical dependent variable depends on the frequency distribution of that variable. For a dichotomous dependent variable, for instance, variance is at a maximum for a 50-50 split and the more lopsided the split, the lower the variance. This means that R-squared measures for LR models with differing marginal distributions of their respective dependent variables cannot be compared directly, and comparison of logistic R-squared measures with R^2 from OLS regression is also problematic. Nonetheless, a number of logistic R-squared measures have been proposed.

RL-squared is the proportionate reduction in chi-square and is also the proportionate reduction in the absolute value of the log-likelihood coefficient. RL-

squared shows how much the inclusion of the independent variables in the LR model reduces the badness-of-fit. RL-squared varies from 0 to 1, where 0 indicates the independents have no usefulness in predicting the dependent. RL-squared often underestimates the proportion of variation explained in the underlying continuous (dependent) variable (see DeMaris, 1992: 54).

Cox and Snell's R-Square is an attempt to imitate the interpretation of multiple R-Square, but its maximum can be less than 1, making it interpretation difficult. It is part of SPSS output.

Nagelkerke's R-Square is a further modification of the Cox and Snell coefficient to assure that it can vary from 0 to 1. It is part of SPSS output.

Pseudo-R-square is an Aldrich and Nelson's coefficient that serves as an analog to the squared contingency coefficient, with an interpretation like R-square. Its maximum is less than 1.

2.2 Outputs *of LR*

Classification tables are 2 x 2 tables in the LR output, which tally correct and incorrect estimates. The columns are the two predicted values of the dependent, while the rows are the two observed (actual) values of the dependent. In a perfect model, all cases will be on the diagonal and the overall correct estimation will be 100 percent. If the logistic model has homoscedasticity (not a LR assumption), the percent correct will be approximately the same for both rows. Since this takes the form of a crosstabulation, measures of association (SPSS uses lambda-p and tau-p) may be used in addition to percent correct as a way of summarizing the strength of the table.

Lambda-p is a proportional reduction in error (PRE) measure, corresponding to the ratio of (errors without the model - errors with the model)/(errors without the model). If lambda-p is .80, the LR model will reduce our errors in classifying the dependent by 80 percent compared to classifying the dependent by always guessing. Lambda-p is an

adjustment to classic lambda to assure that the coefficient will be positive when the model helps and negative when, as is possible, the model actually leads to worse predictions than simple guessing based on the most frequent class. Lambda-p varies from 1 to (1 - N), where N is the number of cases. Lambda-p = (f - e)/f, where f is the smallest row frequency (smallest row marginal in the classification table) and e is the number of errors (the 1,0 and 0,1 cells in the classification table).

Tau-p is an alternative measure of association. When the classification table has equal marginal distributions, tau-p varies from -1 to +1. Negative values mean that the logistic model does worse than expected by chance. Tau-p can be lower than lambda-p because it penalizes proportional reduction in error for non-random distribution of errors.

2.3 *Assumptions of LR*

LR enables the researcher to overcome many of the restrictive assumptions of OLS regression for the following reasons:

1. LR does not assume a linear relationship between the dependents and the independents. It handles nonlinear effects even when exponential and polynomial terms are not explicitly added as additional independents.
2. The dependent variables neither need to be normally distributed not to be homoscedastic for each level of the independent(s).
3. Normally distributed error terms are not assumed.
4. LR does not require that the independents be interval or unbounded.

However, the following assumptions of OLS regression still apply:

1. *Inclusion of all relevant variables in the regression model*. If relevant variables are omitted, the common variance they share with included variables may be wrongly attributed to those variables, or the error term may be inflated.
2. *Exclusion of all irrelevant variables*. If causally irrelevant variables are included in the model, the common variance they share with included variables may be wrongly

attributed to the irrelevant variables. The more the correlation of the irrelevant variable(s) with other independents, the greater the standard errors of the regression coefficients for these independents.

3. *Error terms are assumed to be independent*. Violations of this assumption can have serious effects. Violations are apt to occur, for instance, in correlated samples, such as before-after or matched-pairs studies.
4. *Linearity*. LR does not require linear relationships between the independents and the dependent, as does OLS regression, but it does assume a linear relationship between the logit of the independents and the dependent.
5. *Additivity*. Like OLS regression, LR does not account for interaction effects except when interaction terms (usually products of standardized independents) are created as additional variables in the analysis.
6. *Independents are not linear functions of each other*. To the extent that one independent is a linear function of another independent, the problem of multicollinearity will occur in LR, as it does in OLS regression. As the independents increase in correlation with each other, the standard errors of the logit (effect) coefficients will become inflated. Multicollinearity does not change the estimates of the coefficients, only their reliability.

3. *Cluster Analysis*

Overview

Cluster Analysis (CA) is a multivariate analysis seeking to capture most of the information contained in the original variables so that relatively homogeneous groups, or "clusters," can be formed. The clusters formed should be highly internally homogenous (members are similar to one another) and highly externally heterogeneous (members are *not* like members of other clusters).

The primary objective of CA is to partition a set of objects (e.g., households, individuals) into groups based on the similarity of the objects' characteristics. In forming groups of similar objects the researcher provides a taxonomic description of the data, that is a hypotheses related to the classificatory structure of the data. Also a data simplification is obtained by a CA since objects in the same clusters are easily profiled by general characteristics (such as mean, mode, etc.).

CA has strong mathematical properties and does not need requirements of normality, linearity, and homoscedasticity that are so important in other techniques (e.g., linear regression). CA can accept a wide variety of data and the usual units by variables matrix is substituted by a proximity matrix. The matrix is obtained through computing a proximity measures also be known as "(dis)similarity," "resemblance," or "association." Standardized data are necessary, since otherwise the clustering procedure may be clustering items mainly on larger scales of the considered variables.

The following steps are needed to run a CA:

- selection of the variables and generation of a proximity matrix
- use of a clustering algorithm
- decision about number of clusters and interpretation
- validation of cluster solution

Most commonly used clustering algorithms can be divided into two general categories: (a) *hierarchical procedures* and (b) *nonhierarchical procedures*.

3.1. *Hierarchical Clustering*

In *hierarchical* (*agglomerative*) *procedures*, each of *n* objects is considered as a single cluster. In each subsequent step, close objects are combined into new aggregate clusters, thus reducing the number of clusters. After *n*-1 steps of aggregation the *n* objects are aggregate into a unique cluster. Therefore, joining existing clusters aggregated in previous stages reduces the number of clusters.

A graphical representation of the results of a hierarchical procedure is the *dendrogram* that is a graphical representation depicting the formation of the clusters in the *n*-1 steps. The dendrogram is depicted in a form of a rooted tree (a connected graph without cycles), where the root node represents the set of objects to be classified, while internal nodes of the tree represent clusters obtained at different level of aggregation. Close internal nodes identify close clusters while large distances between internal nodes indicate the presence of distinct clusters.

Hierarchical CA can be applied for relatively small data sets generally not larger than 200-300 objects. In our analysis we used hierarchical CA, and in particular the average linkage method applied on the centroids of a large number of clusters of the entire population determined by the *k*-means algorithm to decide the real number of clusters are present in the data set.

The average linkage method starts out with *n* clusters than aggregate the closest two clusters according to the distance between clusters that at the beginning is the square Euclidean distance. Then the distance between the new aggregated cluster and the remaining *n*-1 is updated. This procedure is repeated for *n*-1 steps of aggregation until all the objects are in a cluster.

3.2. Nonhierarchical Clustering

In contrast to hierarchical methods of CA, nonhierarchical procedures do not

produce the treelike definition of *n*-1 partitions. Instead, it is first necessary to specify a number *k* of clusters that the procedure has to form through an iterative relocation of the objects till the best *k*-clusters solution is reached. The procedure begins by randomly or purposely selecting a given number of objects or 'clusters seeds'. Then, the program aggregates each object to the nearest seed. Then the cluster seeds are revised (updated) by calculating seed cluster means each time the objects are assigned. The procedure stops when no object is moved into a different cluster between two iterative relocation steps. Nonhierarchical procedures are faster than the hierarchical ones. Algorithm such are *k*-means may be applied even for several thousand of objects.

3.3. Strategy of the adopted analysis

The objective in our analysis is to find groups of households that are similar according to a set of variables associated with different aspects of welfare. The derived clusters reflect the inherent structure of the data only as defined by the set of variables describing the welfare of the households. CA technique has no means of differentiating relevant from irrelevant variables, since it detects groups of objects across all variables.

Therefore, to define clusters that are homogeneous for welfare it is first necessary to capture the relevant information of many welfare characteristics in a small number of dimensions. These dimensions, created by the NLPCA, are composite standardized and optimally scale measures of the original variables. The standardization and optimal scaling of the variables is very important for a successful CA, since distances are very sensitive to different scales or magnitude among variables. For this reason, the CA uses the dimensions of the NLPCA instead of the original variables.

Since CA has to determine groups of similar objects, it needs to prefix the measure to evaluate the similarity or dissimilarity between object to be clustered. The concept of dissimilarity is equivalent to distance such as the Squared Euclidean Distance that has been used in our analysis.

The number of clusters in the *k*-means algorithm must be defined a priori. To decide the best number of cluster to be retained the *k*-means algorithm is applied fixing a large number of clusters (k=10). The centers of the clusters, called centroids, are represented by the average profile of the objects in the clusters. A hierarchical CA is applied (average linkage method) on the centroids to see if they are close and therefore associated to other clusters.

The analysis of variance (ANOVA) can be used to test the null hypothesis of no cluster differences. The within cluster estimate of variance and the between clusters estimate of variance represent independent estimates of the population variance. Therefore, the ratio of the two variances is a measure of how much variance is attributable to different clusters versus the variance expected from the population. Larger is the value of this *F* statistics and higher is the probability that the clusters are really different. Although the F test is valid only if normal and equal variance assumptions are verified, there is evidence that *F* tests in ANOVA are robust with regard to these assumptions. For this reason we used this analysis to validate the choice of the number of clusters in the clustering results.

4. *Nonlinear Principal Component Analysis*

Overview

The Nonlinear Principal Component Analysis (NLPCA) is performed using the procedure PRINCALS (SPSS ver. 8.0). This acronym comes from Principal Component Analysis and Alternating Least Squares. Given a set of nominal and categorical variables PRINCALS has the aim to obtain optimal quantifications of the categories. A quantification of categories is optimal if it enhances the properties of the observed data. The quantifications of the original variables are modified by PRINCALS to obtain quantifications relative to new dimensions. This is obtained through maximizing the eigenvalues (variance) of the dimensions, and accounting for the maximum variance of the original variables.

The quantifications are standardized variables and therefore are comparable. Compared to a Principal component Analysis (PCA) solution, that include only objects with no missing data, NLPCA just ignores the value that missing data may have, including all the case. PRINCALS replaces missing data by quantifications that are as much as possible in agreement with the optimal quantification.

PRINCALS is very flexible, since it allows each individual variable (question) to be treated, as it is, i.e. nominal or ordinal or numerical. If all variables are numerical PRINCALS gives solutions similar to PCA with the advantage to evaluate nonlinear relationships between variables. If variables are treated all as nominal the solution of PRINCALS is similar to the Multiple Correspondence Analysis also called Homogeneity Analysis (HOMALS).

4.1 *Optimal Category quantifications*

The optimal category quantifications are obtained by multiplying the initial given category quantification by the *component loadings*. Since category quantifications define

variables with unit variance, the optimal quantifications have variance equal to the square of the component loadings. These values are called the *discrimination measures* or *fit per variable and dimension*.

The sum of the discrimination measures over dimensions is called *fit per variable*. The fit per variable cannot be larger than the number of dimensions in the solution.
If all the *m* initial variables to be optimally quantified are treated as single (numerical, ordinal, nominal), PRINCALS can produce *m* possible dimensions.
If all variables are treated as multiple nominal, and if there are no missing data, PRINCALS solution will be the same as HOMALS (multiple correspondence analysis), and supposing *j*-th variable has k_j categories, the total number of dimensions will be equal to $\sum k_j - m$.

In the mixed case where single and multiple nominal variables are observed, the possible solutions are between *m* and $\sum k_j - m$.