SOME CONSIDERATIONS RELATED TO POVERTY USING THE PRINCIPAL COMPONENTS ANALYSIS (PCA)

Evgjeni Xhafaj[1], Ines Nurja[2]

[1]Department of Mathematics, Faculty of Information Technology, University of Durres, Albania, E mail: genaxhafa@yahoo.com
[2]University of New York Tirane, Albania, E mail:inesheba@gmail.com

Abstract

The measurement and analysis of poverty have traditionally relied on reported income or consumption and expenditure as the preferred indicators of poverty and living standards. Income is generally the measure of choice in developed countries but a number of methods have been used to assess poverty levels and trends which rely not on consumption or income data but rather on non-monetary dimensions of living conditions. The purpose of this study is to make an estimation of the poverty level not based in the incomes but identifying the individual factors which mostly affect poverty and living conditions. In particular, it analyzes several aspects of poverty by using the Principal Components Analysis (PCA). The data used for the analysis in this paper come from Living Standards Measurements Surveys (LSMS) in 2008. The principal components analysis was used to create an asset index which gave the social economic status of each household. Furthermore we also propose an analysis of deprivations and restrictions of Albanian households by analyzing characteristics of housing, utilities and durable ownerships, For the deprivation analysis two different sets of variables are considered. The first set concerns with the possession of certain goods (such as CD player, washing machine, video recorder, fridge, etc). A second set of items relates to basic housing furniture and conditions. The questions investigate if individuals possess some specific item or service such as heating, drinking water, internal toilet or kitchen.

Keywords: *lsms, poverty, principal component analysis, welfare, household*

Introduction:

The measurement and analysis of poverty have traditionally relied on reported income or consumption and expenditure as the preferred indicators of poverty and welfare.

Income is generally the measure of choice in developed countries while the preferred metric in developing countries is an aggregate of a household's consumption expenditures,(Sahn and Stifel 2003). The choice of expenditures over income is influenced by the difficulties involved in the measuring income in the developing countries. Similarly with the expenditure

data the limitation is the extensive data collection which is time- consuming and costly as stated by(Vyas and Kumaranayake 2006).

There are many reasons to measure welfare, which are not based in the expenditure of consumption but in the so called non monetary poverty. The non monetary poverty consists of indicators that don't have a relation with the monetary aspect but with the access to the base services and their quality. Often, to fill in the monetary aspect of poverty with non monetary one, is used a kind of index which is called the base unfulfilled needs index. This index consists of five indicators that includes: the inadequacy water and toilet supply (the absence of water and WC in the buildings), the inadequacy of shelter conditions (according to the household perception), inadequacy of power supply, overpopulation of the residence (three or more individuals per room) and the inadequacy of educational level of the household head (primary or lower level). A household is considered poor if two or more base needs from the base needs are not fulfilled and extremely poor if three of these are not fulfilled (instat).

Filmer and Pritchett (2001) used Demographic and Healthy Survey data to show that the relationship between wealth and enrollment in school can be estimated without income or expenditure data, by using household asset variables. PCA provided acceptable and reliable weights for an index of asset to serve as a measure for wealth Filmer and Pritchett (2001) and Sahn and Stifel (2000- 2003) share the approach of Montgomery et al (1999) in trying to assess welfare with an asset-based measure, but depart from them in two fundamental ways. First, they do not consider the asset index to be a proxy, but rather an alternative, to the consumption measure. Secondly, they introduce the use of principal components (Filmer and Pritchett, 2001) or factor analysis (Sahn and Stifel, 2000; 2003) in order to determine the weights to be assigned to each element of the asset index.

Booysen(2002) used demographic and healthy survey to measure differences in socioeconomic status of South Africa households. The asset index used represented a comparable indicator of poverty in South Africa.

In this paper, we construct an asset index using Principal Components Analysis (PCA) from asset ownership variables in the Living Standards Measurement Study (LSMS, 2008).

Methods:

The data used to analyze the poverty is taken from the 2008 Living Standards Measurement Study (LSMS) for Albania. The survey covered both rural and urban populations. The survey collected information relating to demographic and detailed information on asset ownership, concerns with the possession of certain goods and housing characteristics. A household was defined as a person or a group of people related or unrelated to each other, who live together in the same dwelling unit and share a common source of food. The larger samples were employed to identify the best set of variables and their weights, and the smaller samples were used to test out-sample the prediction accuracy of the constructed tools. To compute the poverty index, the PCA procedure involves a number of steps following Henry et al. (2003) that are illustrated using the example of Bangladesh. First of all, bivariate correlation analyses of the expenditures per capita were run with the initial variable list of 90 variables. Forty variables with highly significant coefficients (alpha < 0.001) and a theoretically consistent sign for the correlation coefficient were retained from the initial data set. Second, before applying the PCA, following

Henry et al. (2003), we grouped these forty variables into several dimensions of poverty. Within each dimension, we dropped variables that were redundant, i.e. they exhibited a high correlation with other variables contained in the same dimension. Third, the PCA was then carried out with SPSS. Here, the maximum number of iterations was set at 25. The Eigen value was limited to 1. Since PCA does not provide an easy way to generate a best fit for a poverty index, a trial and error process using the final 31 variables was used to determine which combination yielded the best accuracy performance. Applying these screening procedures leads to increases in the Kaiser-Meyer-Olkin measure of sampling adequacy (KMO). The larger the KMO index, the higher is the fraction of variance explained by the model. As stated by Henry et al. (2003), the higher the coefficient size, the stronger the relation with the derived poverty index.

We applied PCA to create an asset index based on data from the LSMS (2008). The LSMS (2008) included information regarding the ownership of durable goods, housing characteristic. Using PCA, we first recoded the household variables into dichotomous variables. The PCA is a multivariate statistical technique used to reduce the number of variables without losing too much information in the process. The PCA technique achieves this by creating a fewer number of variables which explain most of the variation in the original variables. The new variables which are created are linear combinations of the original variables. The first new variables will account
for as much as possible of the variation in the original data.
Given p variables $X_1\ X_2,...,X_p$ measured in n households, the p principal components $Z_1\ Z_2,...,Z_p$ are uncorrelated linear combinations of the original variable $X_1\ X_2,...,X_p$, given as.

$$Z_1 = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p$$
$$Z_2 = a_{21}X_1 + a_{22}X_2 + ...a_{2p}X_p \qquad (1)$$

$$Z_p = a_{p1}X_1 + a_{p2}X_2 + ...a_{pp}X_p$$

This system of equations can be expressed as $\boldsymbol{z}=\boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{z}=(Z_1\ Z_2,...,Z_p)$ $\boldsymbol{x}=(X_1\ X_2,...,X_p)$ and $\boldsymbol{A}$ is the matrix of coefficients.

The coefficient of the first principal component $a_{11}, a_{12,}...a_{1p}$ are chosen in such a way that the variance of Z_1 is maximized subject to the constraint that $a_{11}^2 + a_{12}^2 + ... + a_{1p}^2 = 1$
The variance of this component is equal to λ_1, the largest eigen value of $\boldsymbol{A}$. The second principal component is completely uncorrelated with the first component and has variance equal to λ_2, the largest eigen value of $\boldsymbol{A}$. This component explains additional but less variation in the original variable than the first component subject to the same constraint. Further, principal components (up to the maximum of p) are defined in a similar way. Each principal component is uncorrelated with all the others and the squares of its coefficients sum to one. The principal component analysis involves finding the eigen values and eigenvectors of the correlation matrix (Basilevsky, 1994).

Principal components index

The first principal component, the linear combination capturing the greatest variation among the set of variables, can be converted into factor scores, which serve as weights for the creation of the marginality index. The first principal component, the linear combination capturing the greatest variation among the set of variables, can be converted into factor scores, which serve as weights for the creation of the marginality index. Formally:

$$A_j = \sum_{g=1}^{G} F_g * \frac{(x_{jg} - u_g)}{\sigma_g} \quad (2)$$

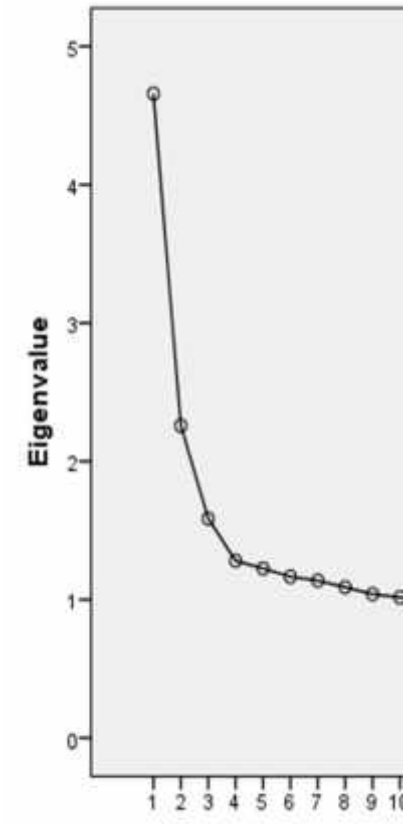
Where, as above, the subscript *g* refers to the asset item, *G* to the total number of different items sampled and *j* to the household. The term F_g represents the variable weights, i.e. the scoring coefficients of each factor's eigenvector, x_{jg} is the value of the g-*th* variable for the *jth* household, and μ and σ are, respectively, the mean and the standard deviation of the *g-th* variable over all households. In principal components, the eigenvector provides the factor score for each variable, which indicates the direction and weight of the impact of each variable in the poverty index.

Results

The results of PCA indicate that the first principal component explains 14.5% of the variation in the original variables and each subsequent component explains a decreasing proportion of variance. The results from principal components analysis can be found in Table 1 where is submitted the 31 Eigen values of the correlation matrix. The eigenvector associated with the first component can be found in Table 2. In principal components, the eigenvector provides the factor score for each variable, which indicates the direction and weight of the impact of each variable in the poverty index. The signs on all variables are as expected. The Kaiser-Meyer-Olkin KMO) of sampling adequacy is relatively high 0,8. The bigger the KMO index, the higher is the fraction of variance explained by the model. The scree plot in Figure 1 shows the proportion of variance explained by each principal component and indicates that the first ten components would sufficiently explain the original variables.

Table 1.

Componet	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4,659	14,559	14,559	4,659	14,559	14,559
2	2,258	7,055	21,614	2,258	7,055	21,614
3	1,586	4,955	26,569	1,586	4,955	26,569
4	1,281	4,003	30,572	1,281	4,003	30,572
5	1,225	3,828	34,400	1,225	3,828	34,400
6	1,166	3,642	38,042	1,166	3,642	38,042
7	1,137	3,553	41,595	1,137	3,553	41,595
8	1,090	3,405	45,000	1,090	3,405	45,000
9	1,040	3,249	48,249	1,040	3,249	48,249



10	1,016	3,175	51,424	1,016	3,175	51,424
11	,995	3,110	54,534			
12	,953	2,977	57,511			
13	,933	2,917	60,428			
14	,901	2,816	63,244			
15	,887	2,773	66,017			
16	,868	2,711	68,728			
17	,844	2,637	71,365			
18	,818	2,557	73,922			
19	,754	2,358	76,279			
20	,739	2,309	78,588			
21	,720	2,249	80,838			
22	,703	2,198	83,036			
23	,660	2,063	85,099			
24	,620	1,937	87,036			
25	,613	1,916	88,952			
26	,593	1,854	90,806			
27	,583	1,823	92,629			
28	,546	1,705	94,334			
29	,517	1,615	95,949			
30	,498	1,557	97,506			
31	,469	1,467	98,973			

Table 2.

	First Eigenvector
Bicycle	,038
Video camera	,105
Car	,103
Color TV	,113
Computer	,110
Conditioner	,120
Electric orgasstove	,114
Microwave	,105
Video DVD	,114
Continuous water	,002
Refrigerator	,072
Satellite	,097
Time of construction	,038
Radiator electric	,036
Distance from: primary school	-,069
Distance from: Ambulatory	-,068

Conclusion:

PCA exploit the association among the variables to identify the latent factors (components) which are weighted function of the original variables; the weights are based on the correlations between the variables and, hence, are optimally defined. A higher association among the variables is an indicator that, these variables capture the same dimension of deprivation. The variables with higher component loadings prevail in the explanation of the deprivation dimension represented by each factor. A factor score is assigned to each individual and it is interpreted as an indicator of the individual condition on each dimension of deprivation. Higher values of factor scores allow identifying the people in various dimensions of deprivation.

The first component was always the one that was identified as the multidimensional poverty index based on a number of criteria. This is because the poverty component and its significant underlying indicators can be identified by analyzing the signs and size of the indicators in relation to the new component variable. Having a negative value for the poverty index identifies a household as being poorer than the population mean, whereas positive values indicate an above-average wealth.

The larger the KMO index (0,8), the higher is the fraction of variance explained by the model. As stated by Henry et al. (2003), the higher the coefficient size, the stronger the relation with the derived poverty index

Most of the above studies report what appear to be robust estimates of welfare rankings and call for a more extensive use of the asset index in poverty measurement and analysis. The main features of this method that make it attractive are: i) the greater availability of household surveys not collecting detailed consumption expenditure or income data; ii) the greater degree of standardization of these surveys – which may reduce the scope for systematic errors due to questionnaire design or administration, and allow greater cross-country comparability; and iii) the possibility of avoiding price indexes, which constitute an additional potential source of bias in consumption-based measures.

References

Basilevsky, A., 1994. *Statistical factor analysis and related methods,* John Wiley and Sons, New York.

Booyen, R (2002). *Using Demographic and Health Surveys To Measure pov*erty. An Application to South Afric

Henry, C., Sharma, M., Lapenu, C., Zeller, M., 2003. *Microfinance poverty assessment tool*,

Tech. T. S. 5, *Consultative Group to Assist the Poor* (CGAP) and The World Bank, Washington,

Filmer, D. & Pritchett, L. 2001. *Estimating Wealth Effects without Expenditure Data or Tears*: An Application to Educational Enrollments in States of India.

www.instat.gov.al

Sahn, D.E. & Stifler, D. 2003. *Exploring Alternative Measures of Welfare in the Absence of Expenditure Data, Review of Income and Wealt*h.

Vyas, S and Kumaranayake, L (2006). *Constructing social-economic status indices, How to use PCA. Health Policy and Planning*