International Journal of Research in Social Sciences
Vol. 9 Issue 2, February 2019,
ISSN: 2249-2496 Impact Factor: 7.081
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Wealth Index of India using Principal Component Analysis

Uma Rani
Department of Statistics, BSA College, Mathura, India
Email- cschaudhary1978@gmail.com

Abstract
Commonly available survey data for developing countries often do not include income or expenditure data. This data limitation puts severe constraints on standard poverty and inequality analyses. We provide a simple approach to simulate household income based on publicly available Demographic and Health Surveys (DHS) and macroeconomic data. We illustrate our approach with India. This paper presents the calculation of wealth index using standard of living index score based on household's assets suggested by DHS and used principal component analysis (PCA) to reduce the dimension of variables. At the end, both approaches have been compared and conclusion has been drawn.

Keywords - Wealth Index, Standard of Living Index, Principal Component Analysis

Introduction

Household income or expenditure data are often used to measure current and long-term welfare of households and within-country inequality. The availability of household survey data has increased the understanding of within-country inequality and its determinants. Large scale national representative household survey data has become more and more available in recent years. However, oftentimes commonly available survey data for developing countries – such as the Demographic Health Surveys (DHS) – do not include income or expenditure data.

Filmer and Pritchett (2001) and Sahn and Stifel (2001) have proposed a one-dimensional index based on household assets and other household characteristics as a proxy of long-term material welfare to overcome the problem of missing income and expenditure data. The so-called 'asset index' is often used in the empirical literature on poverty and inequality analysis as a proxy variable for household income. There is a large body of literature that uses an asset index to explain inequalities in educational outcomes (e.g Bicego et al(2003); Ainsworth and Filmer, 2006), health outcomes (e.g. Bollen et al., 2002; Schellenberg et al., 2003), child malnutrition (e.g. Sahn and Stifel, 2003; Tarozzi and Mahajan, 2005), or child mortality (e.g. Sastry, 2004) when data on income or expenditure is not available. In addition, asset indices are used to analyze changes and determinants of poverty (Harttgen and Misselhorn, 2007; Sahn and Stifel, 2000; Stifel and Christiaensen, 2007; World Bank, 2006). Although the asset index has some shortcomings, which we will address below, it has become a popular tool to overcome the problem of missing data on income or expenditure.

This paper deals with the approach of calculation of wealth index based on thirty two variables suggested by DHS using SLI score then principal component analysis (PCA) has been used for the dimension reduction. Then after again the wealth index has been calculated based on the new obtain variables after the dimension reduction and result has been compared.

Data Description

This paper used a dataset of household's which is obtain from Demographic and Health Survey (DHS) India name household's dataset 2005-06. These variables are given in Table-1which are taken from the dataset of households collected by National family health Survey (NFHS-3) on Indian households based on questionnaire of households.

Statistical Tools Used

i. Standard of living index (SLI)

Weighting of indices

Out of necessity, many standard of living indices are often composed of indirect or proxy indicators rather than direct measures. It is, therefore, unsurprising that a bewildering array of indices has been proposed, using different combinations of variables and different statistical methods.

Methods of weighting indices

Where researchers are very experienced, it is often possible for them to produce very good pragmatic weightings based on a lifetime of research experience. However, few people have this level of in-depth knowledge and most have to rely on more formal methods. There are some general and proven methods of weighting indices that have been developed by European researchers (particularly Dutch, Swedish, Irish, Portuguese and British social scientists).

- Possession weighting
- Proportionate possession weighting
- Opinion weighting
- Proportionate opinion weighting

Possession weighting was suggested by Peter Townsend (1979) in his study of Poverty in the United Kingdom. It involves measuring the normal level of possession for standard of living or health measures and then weighting each component of an index by this level (or its inverse). For example, if 90% of all households can manage to obtain a school education for their children aged 6-14 and 98% of all households do not need to engage in begging, then these two components of a deprivation index could be given a weighting of 90 and 98, respectively. Those households with the highest score on this index would be the most deprived (poorest). Alternatively, if the purpose is to construct a standard of living index rather than a deprivation index, then the same two items could be given a weighting of 10 (100-90=10) and 2 (100-98=2), respectively. Those households with the highest score on this index would be the wealthiest (richest). The possession weighting method has been widely used by European social scientists, particularly when com- paring survey results from different countries or from different years in the same country.

Proportionate possession weighting (PPW) is an adjustment that reflects the differences between various social and demographic groups. Thus, the PPW approach takes account of these differences by adjusting the weighting for each index item according to significant differences within the population. Account could be taken of the variation in the preferences of any number of different social or demographic groups, such as - sex, age, family composition (whether they are single or couples with or without children) or rurality. Proportionate index weights would

allow for differences in, for example, the different levels of possession of electricity and bullock carts between urban and rural households. Opinion weighting has been used widely in both poverty and health research. For example, Joanna Mack and Stewart Lansley conducted social surveys in the UK to determine the populations views on what were the necessities of life. They then produced deprivation indices that were weighted by public opinion and thus went further than any of their predecessors in an effort to relate the definition of poverty to the view of public opinion and to reduce the impact of arbitrary decisions.

Proportionate opinion indices have not yet been used to any great extent in health research but have been used more widely in the study of poverty and deprivation. Such indices have been used extensively in Scandinavian research where public opinion on the minimum acceptable standard of living has been used to produce indices that are weighted to reflect those differences by sex, age and family composition (whether they are single or couples with or without children). This method of weighting to measure poverty has been called the proportional deprivation index (PDI).

ii. Principal component analysis (PCA)

PCA is a multivariate statistical technique used to reduce the number of variables in a data set into a smaller number of 'dimensions'. In mathematical terms, from an initial set of n correlated variables, PCA creates uncorrelated indices or components, where each component is a linear weighted combination of the initial variables. For example, from a set of variables X_1 through to X_n,

$$PC_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1n}X_n$$

$$PC_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2n}X_n$$
$$\vdots$$
$$\vdots$$
$$PC_m = a_{m1}X_1 + a_{m2}X_2 + \cdots + a_{mn}X_n$$

Where, a_{mn} represent the weight for the m^{th} principal component and the n^{th} variable. The weights for each principal component are given by the eigenvectors of the correlation matrix, or if the original data were standardized, the co-variance matrix. The variance (λ) for each principal component is given by the eigenvalue of the corresponding eigenvector. The components are ordered so that the first component (PC_1) explains the largest possible amount of variation in the original data, subject to the constraint that the sum of the squared weights $(a_{11}^2 + a_{12}^2 + \cdots + a_{1n}^2)$ is equal to one. As the sum of the eigenvalues equals the number of variables in the initial data set, the proportion of the total variation in the original data set accounted by each principal component is given by λ_i/n. The second component (PC_2) is completely uncorrelated with the first component, and explains additional but less variation than the first component, subject to the same constraint. Subsequent components are uncorrelated

Vol. 9 Issue 2, February 2019,
ISSN: 2249-2496 Impact Factor: 7.081
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

with previous components; therefore, each component captures an additional dimension in the data, while explaining smaller and smaller proportions of the variation of the original variables. The higher the degree of correlation among the original variables in the data, the fewer components required to capture common information. Principal components are linear combinations of random or statistical variables which have special properties in terms of variances. For example, the first principal component is the normalized linear combination (the sum of squares of the coefficients being one) with maximum variance. In effect, transforming the original vector variable to the vector of principal components amounts to a rotation of coordinate axes to a new coordinate system that has inherent statistical properties. This choosing of a coordinate system is to be contrasted with the many problems treated previously where the coordinate system is irrelevant. The principal components turn out to bathe characteristic vectors of the covariance matrix. Thus the study of principal components can be considered as putting into statistical terms the usual developments of characteristic roots and vectors (for positive semi definite matrices).

Methodology

i. Methodology for calculating SLI

Firstly, recoded the variables according to thevalues assigned in Table-1. Standard of living indices is a scoring index so the proper scores for every variables have been given according to the SLI score given in Table-1 and simply add up them. Then the final SLI scores have been obtained for each household.

ii. Methodology for constructing wealth index

Information on the wealth index is based on data collected in the household questionnaire. This questionnaire includes questions concerning the household's ownership of a number of consumer items such as a television and car; type of drinking water source; toilet facilities; and other characteristics that are related to wealth status. Each household asset for which information is collected is assigned a weight or factor score generated through principal components analysis. The resulting asset scores are standardized in relation to a standard normal distribution with a mean of zero and a standard deviation of one. As the main objective of this paper is to test that wealth index in case of thirty variables and in case of fifteen variables are same or not. So after calculating wealth index based on thirty and fifteen variables, households have been arranged by wealth score (in ascending order) and then break wealth scores in three parts in both cases and proportion (percentage) has been calculated for all parts. After that the proportion (percentage)in case of thirty variables has been compared to the proportion (percentage) in case of fifteen variables for each class of living standard.

Table-2- Variable of household assets, their values and SLI score

Variable	Value Assigned	SLIScore	Variable	Value Assigned	SLIScore
Housetype	SHNFHS2=3	4	Tractor	SH47W=1	4
	SHNFHS2=2	2	Car	HV212=1	4
Toilet facility/Shared	HV205/ HV225 = 11-15/0	4	Motorcycle/scooter	HV211= 1	3
	HV205/ HV225 = 11-15/1	2	Telephone(mobile or landline)	HV243Aor HV221=1	3
	HV205/HV225 =21-23/0	2	Refrigerator	HV209= 1	3
	HV205/HV225=21-23/1	1	Color TV	SH47J=1	3
Electricity	HV206 =1	2	Bicycle	HV210= 1	2
Cooking fuel	HV226=1,2,4	2	Electric fan	SH47G= 1	2
	HV226=5,6,7	1	Radio/transistor	HV207= 1	2
Drinking water source	HV201 =11-12	2	Sewing machine	SH47K=1	2
	HV201=13-3	1	Black and whiteTV	SH47I=1	2
Separate room for cooking	HV242=1	1	Water pump	SH47U=1	2
Own house	SH58=1	2	Animal-drawn cart	HV243C=1	2
Own agricultural land	SH60 =5-990	4	Thresher	SH47V=1	2
	SH60 =2-4.9	3	Mattress	SH47B= 1	1
	SH60 =0-2,999.8	1	Pressure cooker	SH47C= 1	1
Any irrigated land	SH61=0.0-994.0,999.8	2	Chair	SH47D= 1	1
Any live stock	HV246=1	2	Cot/bed	SH47E=1	1

Results

After using principal component analysis, the factor scores of different household assets variable in top six principal component analysis are given in Table-2.These principal components have been further used to calculate the wealth index. Least absolute shrinkage and selection operator (LASSO) method has been used and retain the fifteen variables which has non-zero coefficients. These variables are given in Table-3. The wealth index distribution from thirty variables and fifteen variables are given in Table-4 and Table-5 respectively.

Conclusions

Using principal component analysis, the importance of all the variables has been obtained in the form of weights and using LASSO, fifteen important variables which are useful to calculate the wealth index has been obtained. From Table-4 and Table-5, it can be observed that the wealth

index distributions of all three classes from fifteen variables are not significantly differing from the wealth index distribution from thirty variables. Hence it can be concluded that these fifteen variables can used to calculate the wealth index in place of thirty variables.

Table 2- Factor scores of different variables

Variable	Principal Component					
	1	2	3	4	5	6
Presser cooker	0.753					
Has color TV	0.749		-0.303			
Telephone	0.749					
Toilet facility	0.735					
Table	0.711					
Has refrigerator	0.703		-0.341			
Has chair	0.686					
Has electric fan	0.647					
Has motorcycle/ scooter	0.602					
Mattress	0.592					
Has sewing machine	0.536					
Has kitchen	0.534					
clock/Watch	0.475		0.372			
Type of cooking fuel	0.427				0.361	
Has radio/Transistor	0.375					
cot/Bed	0.364			-0.333		
Areas of agricultural land		0.718			0.403	
Areas of irrigated		0.684			0.366	
land						
Any live stock	-0.354	0.577				
Has animal drawn cart		0.473				
Has water pump	0.303	0.332				
Own house		0.321		-0.316		0.318
Band TV			0.654			
Has car	0.372		-0.39			
Has thresher		0.408		0.483	-0.336	-0.335
Has tractor		0.47		0.474		-0.309
Source of drinking water						
Has electricity	0.313			0.376	0.448	
House type						
Has bicycle		0.326			-0.317	0.522

Table 3- Variables selected from LASSO and their standard notation

Variables	Standard Notations
Housetype	SHNFHS2
Toiletfacility	HV205
Electricity	HV226
Cookingfuel	HV206
Ownagriculturalland	SH58
Motorcycle/scooter	HV211
Telephone(mobileorland-line)	HV243A or HV221
Refrigerator	HV209
Color TV	SH47J
Electricfan	SH47G
Radio/transistor	HV207
Sewingmachine	SH47K
Pressurecooker	SH47C
Chair	SH47D
Table	HV243B

Table 4- Wealth index distribution from thirty variables

Class	Frequency	Percentage	ValidPercentage	CumulativePercentage
Low	24620	22.6	22.6	22.6
Middle	34308	31.5	31.5	54
High	50111	46	46	100
Missing	2	0	0	100
Total	109041	100	100	

Table 5- Wealth index distribution from fifteen variables

Class	Frequency	Percentage	ValidPercentage	CumulativePercentage
Low	26241	24.1	24.1	24.1
Middle	32736	30	30	54.1
High	50033	45.9	45.9	100
Missing	31	0	0	100
Total	109041	100	100	

References

1. Filmer, D., & Pritchett, L. H. (2001). Estimating wealth effects without expenditure data—or tears: an application to educational enrollments in states of India. Demography, 38(1), 115-132.
2. Sahn, D. E., & Stifel, D. C. (2001). Parental Preferences for Nutrition of Boys and Girls: Evidence from Africa. Available at SSRN 457723.
3. Bicego, G., Rutstein, S., & Johnson, K. (2003). Dimensions of the emerging orphan crisis in sub-Saharan Africa. Social science & medicine, 56(6), 1235-1247.
4. Ainsworth, M., & Filmer, D. (2006). Inequalities in children's schooling: AIDS, orphanhood, poverty, and gender. World Development, 34(6), 1099-1128.
5. Bollen, K. A., Glanville, J. L., & Stecklov, G. (2002). Economic status proxies in studies of fertility in developing countries: Does the measure matter? Population Studies, 5
6. Schellenberg, J. A., Victora, C. G., Mushi, A., De Savigny, D., Schellenberg, D., Mshinda, H., & Bryce, J. (2003). Inequities among the very poor: health care for children in rural southern Tanzania. The lancet, 361(9357), 561-566.
7. Sahn, D. E., & Stifel, D. (2003). Exploring alternative measures of welfare in the absence of expenditure data. Review of income and wealth, 49(4), 463-489.
8. Tarozzi, A., & Mahajan, A. (2005). Child Nutrition in India in the Nineties: A story of increased gender inequality?. Available at SSRN 730526.
9. Sastry, N. (2004). Trends in socioeconomic inequalities in mortality in developing countries: the case of child survival in Sao Paulo, Brazil. Demography, 41(3), 443-464.
10. Misselhorn, M., Klasen, S., Harttgen, K., & Grimm, M. (2007). A Human development index by income groups.
11. Sahn, D. E., & Stifel, D. C. (2000). Poverty comparisons over time and across countries in Africa. World development, 28(12), 2123-2155.
12. Stifel, D., & Christiaensen, L. (2007). Tracking poverty over time in the absence of comparable consumption data. The World Bank Economic Review, 21(2), 317-341.
13. World Bank. (2006). World development report 2007: Development and the next generation. Washington, DC: The World Bank and Oxford University Pre.