Research Methods

Assets as a Socioeconomic Status Index: Categorical Principal Components Analysis *vs.* Latent Class Analysis

Majid Sartipi MD PhD Candidate[1], Saharnaz Nedjat MD PhD•[1,2], Mohammad Ali Mansournia MD PhD[1], Vali Baigi PhD Student[1], Akbar Fotouhi MD PhD[1]

Abstract

Background: Some variables like Socioeconomic Status (SES) cannot be directly measured, instead, so-called 'latent variables' are measured indirectly through calculating tangible items. There are different methods for measuring latent variables such as data reduction methods e.g. Principal Components Analysis (PCA) and Latent Class Analysis (LCA).

Objectives: The purpose of our study was to measure assets index- as a representative of SES- through two methods of Non-Linear PCA (NLPCA) and LCA, and to compare them for choosing the most appropriate model.

Methods: This was a cross sectional study in which 1995 respondents filled the questionnaires about their assets in Tehran. The data were analyzed by SPSS 19 (CATPCA command) and SAS 9.2 (PROC LCA command) to estimate their socioeconomic status. The results were compared based on the Intra-class Correlation Coefficient (ICC).

Results: The 6 derived classes from LCA based on BIC, were highly consistent with the 6 classes from CATPCA (Categorical PCA) (ICC = 0.87, 95%CI: 0.86 – 0.88).

Conclusion: There is no gold standard to measure SES. Therefore, it is not possible to definitely say that a specific method is better than another one. LCA is a complicated method that presents detailed information about latent variables and required one assumption (local independency), while NLPCA is a simple method, which requires more assumptions. Generally, NLPCA seems to be an acceptable method of analysis because of its simplicity and high agreement with LCA.

Keywords: Index, latent class analysis, principal components analysis, socioeconomic status

Cite this article as: Sartipi M, Nedjat S, Mansournia MA, Baigi V, Fotouhi A. Assets as a socioeconomic status index: Categorical principal components analysis *vs.* latent class analysis. *Arch Iran Med.* 2016**; 19(11):** 791 – 796.

Introduction

Socioeconomic status (SES) is defined as the access of a person to financial, social and cultural assets and human capital.[1] Certain phenomena in social sciences, behavioral sciences and health sciences, like SES, cannot be directly measured and studied. However, using models people can be classified into subgroups or categories based on their behaviors or traits. Behaviors and traits, which are not measured directly, are called latent behaviors or traits. As seen in Figure 1, a latent variable is generally measured using indicator variables. Today, assets are increasingly used as manifest variables, especially in developing countries, to obtain the socioeconomic status of populations when income and expenditure data are not available or are difficult to collect.[2]

In recent years, data reduction methods have been turned into typical methods for combining various asset variables, particularly in SES studies. To this end, different statistical methods have been used, which are selected upon considering the type of data. Currently, Principal Components Analysis (PCA) is one of the best solutions for achieving data reduction of continuous data.[3]

Figure 1 shows a method for dividing the approaches to analysis, which was performed, based on the type of latent and indicator variables.[4]

According to the data, different methods are used to analyze latent models. When latent variables and indicators are continuous, factor analysis (FA) is employed. If both variables are discrete, latent class analysis (LCA) is applied. The third condition is when the latent variables are continuous and the indicators are discrete; in this case the appropriate analyses to apply are Latent Trait Analysis or Item Response Theory. If latent variables are discrete and indicators are continuous, the appropriate analysis to employ would be Latent Profile Analysis.[4] When the dataset contains variables with different measurement levels, another method called Nonlinear Principal Components Analysis (NLPCA) is used.[5] In addition to the above mentioned options, many studies use the PCA method for the sake of convenience regardless of the type of variables, and the limitations of this method in analyzing categorical, ordinal and binary variables.[6]

LCA divides the community into paired incompatible classes. NLPCA can be performed in SPSS using the CATPCA command. Despite its advantages, including its helpful information on latent variables, LCA is not generally used due to its complexity in writing syntaxes, as well as the need for advanced software.

The aim of NLPCA and PCA is to reduce variable numbers to a smaller number of non-correlated principal components, to preserve the information embedded in the data. This goal may be

Authors' affiliations: [1]Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran, [2]Knowledge Utilization Research Center, Tehran University of Medical Science, Tehran, Iran.
•Corresponding author and reprints: Saharnaz Nedjat MD PhD, Knowledge Utilization Research Center, Tehran University of Medical Science, Tehran, Iran. Cell Phone: +98-912-1406265, Fax: +98-21-8888-9123, E-mail: nejatsan@tums.ac.ir.
Accepted for publication: 5 October 2016

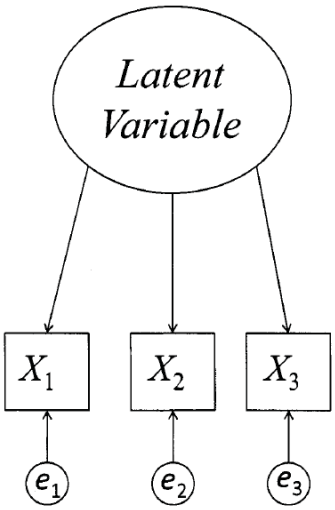

Figure 1. Latent variable with 3 indicator variables[4]

achieved by finding small numbers of linear combinations which explain the maximum variances of data. The main advantages of NLPCA versus PCA are as follows: 1) NLPCA can be used in non-linear relationships; 2) it enables researchers to study multivariate linearity between variables readily, especially in ordinal data; and, 3) it is possible to simultaneously analyze variables with different scales. In other words, NLPCA overcomes the general limitations of the linear PCA i.e. the assumption of linearity of relationships.[5]

LCA classifies people into subgroups and separate definite classes in terms of latent variables using indicator variables, which are measured directly.

Factor analysis analyzes a covariance matrix to determine the basic latent structure. In factor analysis, the latent variable's scores form a continuous distribution. For example, while factor analysis creates a continuous distribution of 'skill scores', LCA divides a community into discrete latent classes like 'highly skilled', 'moderately skilled' and 'un-skilled' classes. The only assumption in LCA is the *local independency*; the observed variables are independent, conditional on the latent variable. This model does not need the assumption of normality.[4]

This study aimed to apply both LCA and NLPCA analyses on a dataset including binary indicator variables and to compare the results of these analyses.

Methods

Study population

The study was designed and carried out in Tehran in 2013. One thousand nine hundred ninety-five (1995) persons aged over 18 years were selected through multi-stage sampling of 22 municipalities (as the strata of Tehran). In each municipal area, blocks were randomly selected proportional to size. From each block, 10 families were selected by systematic sampling. From each family, only one person aged over 18 years was selected for face-to face interview through quota sampling -upon considering age and sex criteria. To provide a representative sample, the survey was performed through a multi-stage sampling scheme . In the first step, stratified sampling was established, considering each of the 22 municipality zones of Tehran as strata. Then, cluster sampling was conducted in each stratum, in which 200 blocks were selected through proportional to size. Next, systematic random sampling was performed within selected blocks. Therefore, the interviewer had to count all places of residence except vacant/abandoned houses, non-residential buildings and pensions. Then, the number of places of residence should be divided by ten to determine the sampling interval. In each household, just one respondent was selected for interviewing through quota sampling considering the sex and age of the city population.

The trained interviewers comprised 10 teams with 4 members each, as well as 4 supervisory teams in charge of data collection.

Out of the 2978 households who were offered to participate in the study, 33% refused to participate (response rate = 66.9%). Data were collected through a questionnaire containing items on socioeconomic status, self-rated health, and objective health. The data was used from the first part of this questionnaire. The data on the socioeconomic status was based on the household's ownership of the selected assets which were shown in Table 1.

Statistical Analysis

Nonlinear Principal Components Analysis (NLPCA)

SPSS 19 was used to perform CATPCA. Since outliers can significantly influence analysis results, an object plot was derived for two dimensions to avoid them in the output results of NLPCA. In case they did exist, they were excluded (Figure 2). The number of dimensions was selected based on the 'eigenvalue greater than 1 criterion'. An Eigenvalue is a ratio of the variance of all variables determined by the relevant dimension. In other words, an eigenvalue is the relative contribution of each dimension to the total variance of all variables. Loadings show the correlation of each variable with its relevant dimension.[7] To test the internal consistency of assets data, Cronbach's alpha was used. A Cronbach's alpha greater than 0.7 indicates acceptable internal consistency of that data.[8] This means that the data variance originates from the interviewees' answers, not from the design and use of the questionnaire.[9,10] In the NLPCA method, the first factor accounting for the major part of the variance was conventionally selected as the respondents' economic status.

Table 1. Assets frequencies among respondents (n = 1995)

Assets	Frequency	Percent
Car	1160	58.1
Freezer	1674	83.9
Dish Washing	550	27.6
Microwave Oven	807	40.5
PC	1264	63.4
Vacuumc	1896	95
Washing Machine	1822	91.3
LCD or LED TV	1524	76.4
Video	1189	59.6

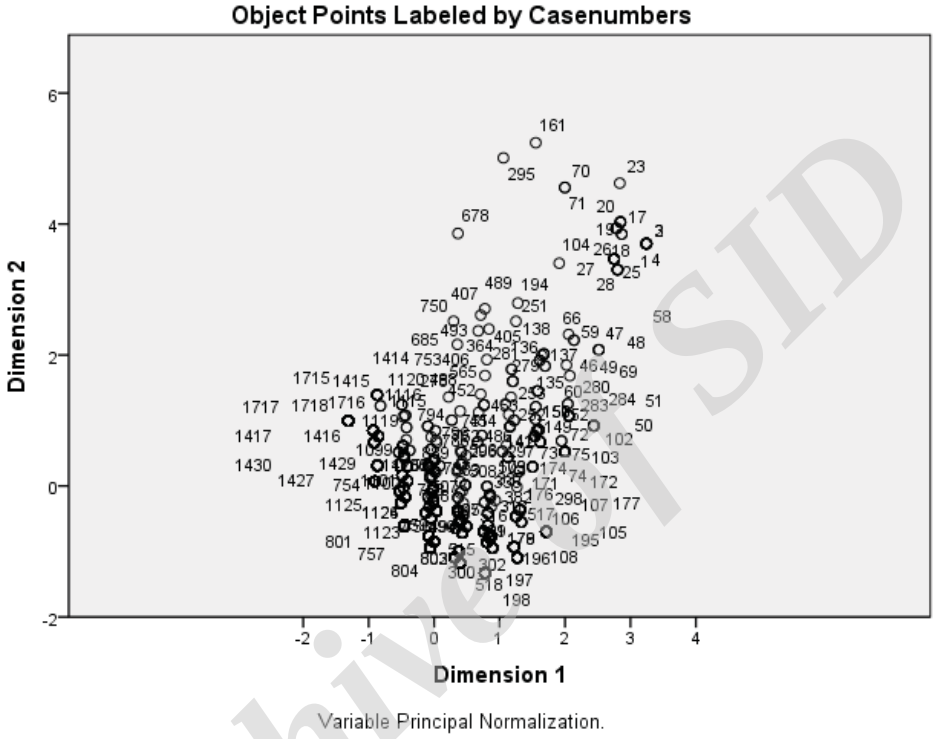

Figure 2. Object plot depicting dimension scores on components 1 and 2

Latent Class Analysis (LCA)

Following the preparation and writing of syntaxes using PROC LCA, LCA analysis was conducted on the data of 1979 respondents in SAS 9.2. Software.[11,12] The model with the minimum values of the AIC and BIC measures was selected as the preferred model. AIC and BIC are criteria that can be used to compare different models. These two criteria are calculated using the following formula:

$$“AIC = G^2 + 2P$$
$$BIC = G^2 + [log(N)]P”$$

Where *P* is the number of estimated parameters and log refers to the natural log. The lower amount of AIC or BIC implies better model.[4] Following LCA analysis, the following 6 classes were ranked in terms of the probability number of assets from richest to poorest (richest, rich, upper-middle, lower-middle, poor and poorest). Using these data, each respondent was assigned to the class with the *best* probability. For instance, respondent number 1018 in class no. 6 had a probability of 0.947, while the same respondent had a probability of 0.053 in class no. 5, and a probability of zero in other classes. Therefore, this respondent was classified in class 6 i.e. the poorest class. This procedure was repeated for each respondent.

Comparison of NLPCA and LCA

To compare NLPCA and LCA, dimension 1 of the NLPCA model was divided into 6 equal ordinal categories[13] because the LCA model used 6 classes and we needed to obtain the consistency between these two models. Then, the intra-class correlation coefficient (ICC) of the SES variable from both the LCA and NLPCA models was calculated in SPSS.

Results

Nine hundred sixty-three (963) of the 1995 respondents (48.3%) were female (mean age, 41.7 years) and 1032 (51.7%) were male (mean age, 41.9 years). Moreover, 693 respondents (34.9%) held academic degrees, while 1291 (64.7%) respondents held no academic degree; 4 respondents (0.2%) had a seminary

Table 2. Model Summary of CATPCA

Dimension	Cronbach's Alpha	Variance Accounted For	
		Total (Eigenvalue)	% of Variance
1	0.75	2.98	33.08
2	0.16	1.16	12.91
Total	**0.91[a]**	**4.14**	**45.99**

[a]Total Cronbach's Alpha is based on the total Eigenvalue.

Table 3. Component Loadings

	Dimension	
	1	2
Car	0.58	-0.19
Freezer	0.47	0.17
Dishwasher	0.54	-0.48
Microwave oven	0.65	-0.39
PC	0.66	-0.13
Vacuum cleaner	0.48	0.65
Washing machine	0.59	0.51
LCD or LED TV	0.62	0.11
Video	0.55	-0.08

Variable Principal Normalization.

education. One thousand three hundred fifty-six (1356; 68.4%) respondents were married, 488 (24.6%) were single, and 139 (7%) were divorced or widowed. Four member families were the most frequently observed household size, with a percentage of 33.9%, followed by three member families (29%). Only 2.8% of respondents were living alone. Table 1 shows the absolute and relative frequencies of each asset.

One thousand nine hundred seventy-nine (1979) participants were included in the NLPCA analysis. Cronbach's alpha for dimension 1 was 0.75, implying an acceptable internal consistency of data (Table 2).

According to dimension 1 and dimension 2 plots, there was no outlier, therefore, no respondent was excluded (Figure 2). 'Variance Accounted For' (VAF) is a variance explained by the relevant dimension. VAF% is a percentage of total variance by which the dimension is explained. Based on our results, the first two dimensions accounted for 46% of variance scores of all assets. The value of VAF (number of variables multiplied by VAF %) was calculated as $9 \times 0.33 = 2.97$ for the first dimension (Table 2).

In the NLPCA analysis, component loadings were obtained based on two dimensions (Table 3). Tabachnik and Fidell believe that the minimum factor [component] loading required for a variable is 0.32.[12] All loadings of the first dimension were above 0.32 (Min = 0.475 for freezer and max = 0.658 for PC.[7] In the conducted analysis, a score was assigned to each respondent.

The number of required classes for LCA analysis was selected considering BIC.[4] As seen in Table 4, the model with 6 classes had the minimum BIC (2033.14). Moreover, its AIC was relatively low. Less classes are also easier to interpret; the model with 6 classes was selected for analysis. After LCA analysis, the above 9 assets index variables were categorized into 6 classes. Table 5 shows the relative frequency of the possibility of owning each asset by the respondents in their respective classes. For instance, 86% of the respondents from the richest class owned a car, 95% had a freezer, 66% had a washing machine, etc. The 6 aforementioned classes were named arbitrarily. For each class, the mean relative

Table 4. AIC and BIC for selecting number of classes in LCA

Number of Latent Classes	AIC	BIC	G^2	df
1	4596.4	4669.17	4570.40	6130
2	2560.59	2711.72	2506.59	6116
3	2135.35	2364.85	2053.35	6102
4	1759.35	2067.21	1649.35	6088
5	1687.57	2073.79	1549.57	6074
6	1568.55	2033.14	1402.55	6060
7	1546.69	2086.67	1352.69	6046
8	1511.67	2132.98	1289.67	6032
9	1486.17	2185.84	1236.17	6018
10	1472.56	2250.6	1194.56	6004
11	1472.78	2329.19	1166.78	5990
12	1458.59	2420.35	1151.59	5976

Table 5. Latent classes

Title	Richest	Rich	Upper-middle	Lower-middle	Poor	Poorest
Class	1	2	3	4	5	6
Class membership probabilities	0.30	0.37	0.03	0.14	0.14	0.02
Car	0.86	0.69	0.61	0.18	0.19	0.17
Freezer	0.95	0.93	0.81	0.68	0.67	0.18
Dishwasher	0.66	0.14	1	0.02	0	0.07
Microwave Oven	0.92	0.25	0.48	0.14	0.05	0
PC	0.96	0.75	0.38	0.25	0.18	0.11
Vacuum Cleaner	1	1	0.91	0.96	0.91	0
Washing Machine	1	0.99	0.96	0.94	0.69	0
TV	1	0.8	0.74	1	0.1	0.2
Video	0.91	0.61	0	0.49	0.25	0.03
Mean	**0.92**	**0.68**	**0.65**	**0.52**	**0.34**	**0.08**

Table 6. LCA/NLPCA cross tabulation

LCA assets classes	NLPCA Assets classes						Total (%)
	1	2	3	4	5	6	
1	488	76	21	0	0	0	585 (29.7)
2	35	320	151	135	99	0	740 (37.4)
3	5	12	21	10	9	3	60 (3.0)
4	0	13	0	133	84	45	275 (13.9)
5	0	0	0	0	237	33	270 (13.7)
6	0	0	0	0	0	45	45 (2.3)
Total (%)	**528 (26.7)**	**421 (21.3)**	**193 (9.8)**	**278 (14.1)**	**429 (21.7)**	**126 (6.4)**	**1975 (100)**

frequency of assets was calculated in descending order (for example, the means of the first and second classes were 0.92 and 0.68, respectively). The first, second, third, fourth, fifth and sixth ranks were titled the richest, the rich, the upper-middle, the lower-middle, the poor, and the poorest class, respectively based on the mean relative frequency of assets. Then, the likelihood of the presence of each respondent in the above sextet classes was derived using the *Best* syntax of SAS 9.2, i.e. each respondent was assigned to the class with the best probability (Table 5). The second class was composed of 60 respondents, 61.7% of which were living in 6 adjacent municipal areas (2, 3, 4, 5, 8, and 10). Most of them had washing machines (96%), vacuum cleaners (91%) and freezers (91%). Among them, 81.6% were single to four member families (lowly populated), and 61% of the families owned cars. To compare the scores with LCA classes, the scores of the first dimension were ordinarily classified into 6 classes.

Finally, the ICC of the 6 classes of NLPCA analysis and the 6 classes of LCA analysis was calculated and derived as 0.873 (95%CI: 0.862 – 0.883, $P < 0.001$). This ICC indicates that both analyses have a good consistency. Table 6 shows cross tabulation between LCA and NLPCA classes.

Discussion

Both models in determining SES, treated fairly similar. The calculated ICC for the object scores derived from NLPCA and the best possibility of the presence of a respondent in a class were consistent with each other (ICC = 0.87, 95%CI: 0.86 – 0.88, $P < 0.001$).

Based on some evidence, asset is a stable measure of SES and widely uses as a proxy for SES.[14]

Given that the collected data may not be valid on households' incomes and expenditures, their assets were used as an index by which their socioeconomic status could be evaluated.[15] Although assets may not represent SES entirely, they may give an estimate of many respondents' SES. The DHS comparative report describes wealth as the equivalent of net assets that can be theoretically measured.[16] In our two-dimensional NLPCA analysis, all factor loadings were above 0.32 in the first dimension, implying a high correlation between that dimension and its related variable. In other words, all the variables explained more than 10% of variances in this dimension. Therefore, all asset variables remained in the model. The loadings of two variables (vacuum cleaner and washing machine) were above 0.32 in two dimensions, implying the cross-loading of the variables. Since there were a limited number of such variables, they were not excluded from the analysis. Cross-loading variables were correlated with two or more dimensions and their loading value was high (above 0.32) in two or more dimensions. Following analysis, the respondents were categorized into 6 classes based on their object scores. In the LCA analysis, the model with 6 classes was selected by considering its BIC score.

To the best of our knowledge, there is no article in which LCA and NLPCA have been compared. There was only one paper in which 497 adults from Missouri, U.S. had been investigated for 18 ADHD (Attention deficit/ Hyperactivity Disorder) symptoms, wherein the data had been analyzed by LCA and PCA methods. However, this paper has not compared the methods. Rather, it has been compared with a similar study in Brazil (N = 483). The above two studies also indicated that both analytical methods yielded the same results.[17]

In Roskman's study, all 39 psychologists from Nijmegan University's Psychology Department working in 9 separate research studies and training regions were ranked in terms of their work. Each psychologist was considered a variable. First of all, linear PCA analysis was conducted and the variance of the first

two principal components was calculated at 55%. Then, NLPCA was used and the variance of the first two principal components increased by 76.6%.[18] According to the latter study's findings, the NLPCA model is preferred over the PCA when analyzing categorical data.

In the LCA method certain syntaxes need to be written in special software like SAS or 'R'. Moreover, a relatively complex process should be followed to execute this method. This can be a reason it has not been considered by researchers. Factor analysis is a variable-based method emphasizing the identification of relationships between variables. It assumes that these relationships exist among all study respondents. In contrast, LCA is a case-based method wherein subgroups of people with similar characteristics can be identified. This method can accurately divide people into appropriate subgroups based on the studied characteristics. As a subgroup of factor analysis, NLPCA is a simple analysis method that is performed in SPSS, only a few clicks away. The type of model selected depends on our expectations of results. If we are dealing with only one latent variable, the NLPCA seems logical, but if we need more detailed information on the subgroups' characteristics, this model will no longer meet our expectations and we will need more sophisticated models like LCA. In our study, the 6 classes derived from the LCA analysis were ranked from the richest to the poorest. The relative frequency of each asset in the richest and poorest classes was, on average, 0.92 and 0.08, respectively. The second class consisted of relatively rich families from Tehran's affluent regions. They had fewer cars than other rich groups due to their older age. However, they were better off as far as other welfare facilities were concerned, such as having a dishwasher.

Finally, to compare these two methods, the ICC was used for the sake of convenience. Keep in mind that when continuous data are converted into ordinal ranked data, their ICC will remain similar to the weighted kappa.[19]

Study Limitations

One of the main variables in health researches is socio-economic status. Different variables such as, income, household expenditure, job, education, and assets explain this index. However the collected certain data such as expenditure and income may not be valid in many cases. Therefore, we collected data on assets, job, education, number of family members, number of rooms and building area as the variables explaining SES. Among these variables, we only used the assets variable since it was a binary one, to compare LCA and NLPCA models and to reduce the complexity of the models. We recommend researchers to use more completed data, including nominal, ordinal and continuous variables. This way, they may arrive at different ICC results from ours. It should also be noted that there is no gold standard for evaluating socioeconomic status, and the best solution for achieving our study goal is to use simulation, which is recommended for future studies. A simulation is a mathematical business model, which combines both mathematical and logical concepts that tries to emulate a real life system using computer software. Computer simulations use computer models to also predict how a system will behave given a set of conditions. In the case of SES, one can use a sample of data for imitation by a huge iteration to simulate the real SES in the community.

In conclusion, although the LCA and NLPCA methods (for binary data) both presented almost the same results, there were some differences between their findings. Despite the complexity of LCA, it presents more detailed information about latent variables, while NLPCA presents general information. NLPCA requires certain assumptions like normality, while LCA has only one assumption: the independency of indicator variables from the condition of the latent variable.

Since there is no gold standard to measure SES; it is not possible to say definitely that a specific method is better than the other one. However, if somebody wants to use a data reduction method, NLPCA and LCA are likely to show comparable results, despite the fact that LCA can show results in more detailed than NLPCA. This finding can help researchers to choose NLPCA instead LCA, which is a complicated method.

References

1. Cowan CD, Hauser RM, Kominski RA, Levin HM, Lucas SR, Morgan SL, et al. Improving the measurement of socioeconomic status for the national assessment of educational progress: A theoretical foundation. *National center for Education Statistics*, 2003.
2. Filmer D, Scott K. Assessing asset indices. *Demography*. 2012; 49(1): 359 – 392.
3. Kolenikov S, Angeles G. Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer? *Review of Income and Wealth*. 2009; 55(1): 128 – 165.
4. Collins LM, Lanza ST. Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences. 2013; 718: John Wiley & Sons. doi: 10.1002/9780470567333
5. Linting M, van der Kooij A. Nonlinear principal components analysis with CATPCA: A tutorial. *Journal of Personality Assessment*. 2012; 94(1): 12 – 25.
6. Collins M, Dasgupta S, Schapire RE. A generalization of principal components analysis to the exponential family. *In Advances In Neural Information Processing Systems*. 2001; 617 – 624.
7. Osborne JW, Costello AB. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Pan-Pacific Management Review*. 2009; 12(2): 131 – 146.
8. Tavakol M, Dennick R. Making sense of Cronbach's alpha. *International Journal of Medical Education*. 2011; 2: 53 – 55.
9. George D. SPSS for windows step by step: A simple study guide and reference, 17.0 update, 10/e: Pearson Education India,2003 .
10. Gliem, Rosemary R, Gliem JA. Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. *Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education*, 2003.
11. Lanza ST, Linda M. Collins, Lemmon DR, Schafer JL. PROC LCA: A SAS procedure for latent class analysis. *Structural Equation Modeling*. 2007; 14(4): 671 – 694.
12. Tabachnick BG. Clearing Up Your Act: Screening Data Prior to Analysis, Tabachnick, BG & Fidell, LS (eds), Using Multivariate Statistics. 2001.
13. Nedjat S, Hosseinpoor AR, Forouzanfar MH, Golestan B, Majdzadeh R. Decomposing socioeconomic inequality in self–rated health in Tehran. *J Epidemiol Community Health*. 2012; 66: 495 – 500.
14. Howe LD, Galobardes B, Matijasevich A, Gordon D, Johnston D, Onwujekwe O, et al. Measuring socio–economic position for epidemiological studies in low–and middle–income countries: A methods of measurement in epidemiology paper. *Int J Epidemiol*. 2012; 41(3): 871 – 886.
15. Vyas S, Kumaranayake L. Constructing socio–economic status indices: how to use principal components analysis. *Health Policy and Planning*. 2006; 21(6): 459 – 468.
16. Rutstein SO, Johnson K. The DHS wealth asset index. Calverton, Maryland: ORC Macro. 2004(6): 6.
17. Rasmussen ER, Todd ED, Neuman RJ, Heath AC, Reich W, Rohde LA. Comparison of male adolescent-report of attention-deficit/hyperactivity disorder (ADHD) symptoms across two cultures using latent class and principal components analysis. *Journal of Child Psychology and Psychiatry*. 2002; 43(6): 797 – 805.
18. De Leeuw J. Nonlinear principal component analysis and related techniques. *Department of Statistics*, UCLA, 2005.
19. Szklo M, Nieto FJ. Epidemiology: beyond the basics. *Jones & Bartlett Publishers*. 2014; (3): 355.