**Comparison of principal component analysis and multiple correspondence analysis in estimating household wealth index.**

**Objectives**

The objective of this study is to determine the agreement between principal component analysis and multiple correspondence analysis with the related social economic status groups, and to check whether the items are highly correlated to each other i.e to do reliability analysis.

**Method**

Wealth indices were constructed using the assets on which data are collected within Demographic and Health Survey. Construction of indices was done using five weighting methods: Weights equal to the inverse of the proportion of households owning the item, equal weights, PCA, PCA using dichotomised versions of categorical variables, and Multiple correspondence Analysis. Accuracy of the agreement between indices was assed. Indices were compared with per capita consumption expenditure, for household adjust consumption expenditure for household size and composition. Data collected included household assets, household characteristics and utilities. Households were classified into five social economic status and dichotomised into poorest household (poorest 40%), middle (3rd poor 20%) and less poor (richest 40%). Comparison between the accurate agreement was used to calculate percentage of accuracy and Cronbach’s alpha to do reliability test.

**Conceptual framework**

Below is a framework that represents the blueprint of the mathematical model.

**Independent variable/ input**

**Dependent variable**

Household Assets

* Television, radio mobile
* Car, motorbike, bicycle
* Livestock

Household characteristics

* Type of house
* Floor material
* Type of roof
* Occupation of the respondent

Utilities

* Main source of water
* Main cooking fuel
* Type of toilet
* Comparison between PCA and MCA
* Proportion of accuracy or misclassification
* Percentage of accuracy results and misclassification
* Reliability test

**Analysis plan**

KDHS individual record will be analyzed using Stata version14. Descriptive statistics will be performed. I’ll use reliability test to check on the internal consistency between the items on the scale and check the percentage accuracy of the agreement between Principal Component Analysis and Multiple Correspondence Analysis.

The independent variables will be the household characteristics, household assets and utilities for example source of water.

**Result**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 5 quantiles of mca\_csore |  | 5 quantiles of pca\_score | | |  | Total |
|  | 1 | 2 | 3 | 4 | 5 |  |
|  |  |  |  |  |  |  |
| 1 | **1,730** | 1,089 | 1,160 | 1,148 | 1,090 | 6,217 |
|  | 27.83 | 17.52 | 18.66 | 18.47 | 17.53 | 100 |
|  |  |  |  |  |  |  |
| 2 | 925 | **1,016** | 1,031 | 1,377 | 1,866 | 6,215 |
|  | 14.88 | 16.35 | 16.59 | 22.16 | 30.02 | 100 |
|  |  |  |  |  |  |  |
| 3 | 325 | 878 | **930** | 1,245 | 2,838 | 6,216 |
|  | 5.23 | 14.12 | 14.96 | 20.03 | 45.66 | 100 |
|  |  |  |  |  |  |  |
| 4 | 1,102 | 953 | 1,595 | **2**,**151** | 417 | 6,218 |
|  | 17.72 | 15.33 | 25.65 | 34.59 | 6.71 | 100 |
|  |  |  |  |  |  |  |
| 5 | 2,135 | 2,281 | 1,500 | 293 | **4** | 6,213 |
|  | 34.36 | 36.71 | 24.14 | 4.72 | 0.06 | 100 |
|  |  |  |  | |  |  |
| Total | 6,217 | 6,217 | 6,216 6,214 | | 6,215 | 31,079 |
|  | 20 | 20 | 20.00 19.99 | | 20.01 | 100 |

**Analysis and conclusion**

Doing a reliability test to check on the internal coefficient using Cronbach’s alpha, it was highly coefficient which was very good (0.81)

The level between wealth indices and the consumption expenditure did not differ when variety scales of consumption equivalence was applied. Three levels i.e poorest, poor and middle was all equivalent at (20.0%) while the less poor (20.01%) and least poor (19.99%)

On display the accuracy agreement percentage (18.76%) which was lower as to compared to misclassified percentage (81.24%)

On checking the agreement accuracy, we use 70% as the cutoff for high accuracy, since the percentage agreement is (18.76%) we conclude that the accuracy is low between the two models meaning that the two models can give misleading conclusions and should be used with caution.