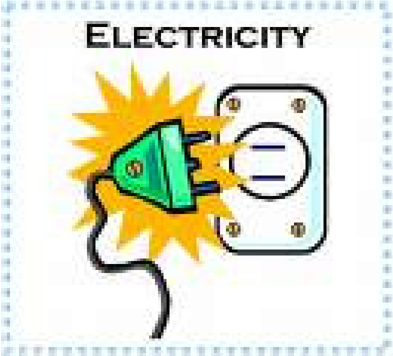
Using Principal Components Analysis to construct a wealth index

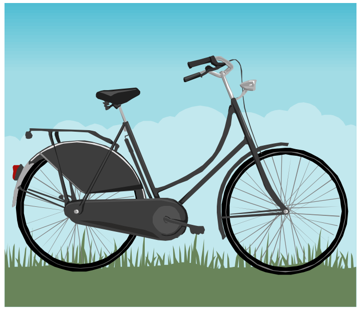
Laura Howe

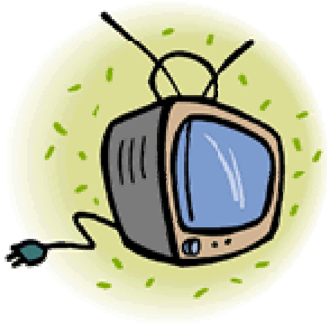
James Hargreaves, Bianca De Stavola, Sharon Huttly

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Wealth Index







Principal Components Analysis

- Data reduction technique
- From set of correlated variables, PCA extracts a set of uncorrelated ‘principal components’
- Each principal component is a weighted linear combination of the original variables

Principal Components Analysis

i.e. if we have *n* correlated variables X_1 - X_n

each principal component is the sum of each variable multiplied by its weight (the weight for each variable is different in each principal component)

$$\mathbf{PC_i=a_1X_1 + a_2X_2 + \ldots + a_nX_n}$$

Principal Components Analysis

- If we use 10 variables in PCA, we get 10 ‘principal components’

- The components are ordered so that the first principal component (PC_1) explains the largest amount of variation in the data

- We assume that this first principal component represents wealth/SEP

Principal Components Analysis

- Assumption:

  The most important reason households have different values of the indicators we have put in the PCA is their wealth/SEP

Issues in using PCA

1. Obscure

2. 1st principal component often explains a low proportion of the total variance

3. **Designed for continuous data**

PCA with discrete data

- Dummy variables most common approach

- ‘Confuses’ PCA:

  Variation arises both from the underlying concept of wealth and from the dependence between dummies for categorical variables

Using dummy variables

- General practice involves omission of the dummy for the baseline category
- In PCA, a dummy is often used for ALL categories
  (Filmer & Pritchett, Vyas & Kumaranayake, DHS report)
- Linear dependence when all dummies are included; affects weights

Alternatives to dummy variables

1. Treat categorical variables as continuous
   REQUIRES ORDINAL VARIABLES

2. Multiple Correspondence Analysis (MCA)

Kolenikov & Angeles, 2004:

- Large simulation study
- Treating ordinal variables as continuous is ‘better’ than using dummy variables

Analysis

Aim:

- Compare the agreement of indices constructed using different methods for dealing with categorical variables

Methods

- Dataset: Malawi IHS2 2004/5, N=11,280
- Wealth index as in DHS of same year
- Agreement measured by classification into quintiles

Methods

- 5 wealth indices:

1. PCA: Include dummies for all categories
2. PCA: Exclude lowest SEP category dummy
3. PCA: Exclude lowest frequency dummy
4. PCA: Treat ordinal variables as continuous
5. MCA

	PCA: dummy for each category	PCA: exclude lowest SEP dummy	PCA: exclude lowest frequency dummy	PCA: treat categorical variables as continuous	MCA
PCA: dummy for each category	-				
PCA: exclude lowest SEP dummy	61 %	-			
PCA: exclude lowest frequency dummy	97 %	60 %	-		
PCA: treat categorical variables as continuous	67 %	42 %	67 %	-	
MCA	76 %	44 %	75 %	87 %	-

	PCA: dummy for each category	PCA: exclude lowest SEP dummy	PCA: exclude lowest frequency dummy	PCA: treat categorical variables as continuous	MCA
PCA: dummy for each category	-				
PCA: exclude lowest SEP dummy	61 %	-			
PCA: exclude lowest frequency dummy	97 %	60 %	-		
PCA: treat categorical variables as continuous	67 %	42 %	67 %	-	
MCA	76 %	44 %	75 %	87 %	-

	PCA: dummy for each category	PCA: exclude lowest SEP dummy	PCA: exclude lowest frequency dummy	PCA: treat categorical variables as continuous	MCA
PCA: dummy for each category	-				
PCA: exclude lowest SEP dummy	61 %	-			
PCA: exclude lowest frequency dummy	97 %	60 %	-		
PCA: treat categorical variables as continuous	67 %	42 %	67 %	-	
MCA	76 %	44 %	75 %	87 %	-

Weights from PCA

- Example 1: floor material

	Index including all dummies
Cement floor	0.6719
Sand floor	0.1823
Tile floor	0.0478
Other floor	0.0209
Mud floor	-0.7160

Including all dummies can result in counter-intuitive weights

Weights from PCA

- Example 2: toilet facility

	Index including all dummies	Index excluding lowest SEP category	Index excluding lowest frequency dummy
Latrine no roof	-0.0611	-0.0631	-0.0659
None/other facility (lowest SEP)	-0.0923	0	-0.1041

Excluding lowest SEP dummy can result in counter-intuitive weights

- Example 3: drinking water source

	Index including all dummies	Index excluding lowest SEP category	Index excluding lowest frequency dummy
Piped into dwelling (lowest frequency)	0.2762	0.3132	0
Piped outside of dwelling	0.1630	0.1663	0.1873

Excluding lowest frequency dummy can result in counter-intuitive weights

- So different methods can alter the order of categories

- Weights from PCA should ‘make sense’

What to do?

- All methods of using dummy variables can be flawed…
- Assign an order to the categories if possible, and treat as continuous
- If not, MCA

Acknowledgements

- Sharon Huttly
- James Hargreaves
- Bianca De Stavola
- Paul Clarke