2011 OPHI-HDCA Summer School on Capability and Multidimensional Poverty

Problem Set on Factor Analysis and Principal Components

by Jose Manuel Roche

Exercise 1 – Examining the implications of generating a 'Wealth Index' using PCA
In this exercise we will explore how the PCA is used as an ad-hoc data reduction technique to generate aggregated indices. We will discuss the advantages and disadvantages of this family of techniques as an aggregation method and as a procedure to set weights.

Data: 'ss_haiti06_Ejercicio1.dta'
Variables to consider in the index:
- Dichotomous variables*:* radio television refrigerator bicycle motorbike car telephone water electricity roof floor
- *Ordinal scales:* cookingfuel toilet wall (the categories of these variable are also avaiable as dummys variables in case you prefer to use these)
- *Continue:* crowding

Sample weight: weight

1. Compute a wealth index using PCA and using the dichotomous variables (this is the method followed by Filmer and Pritchett 2001). If you feel confident, run the analysis with the matrix of polychoric correlations (see commands at the bottom of this document).

   Use the command:
   .factor varlist [if] [in] [weight] [, method options] ➔ use the method pcf

   Save the score from the first component with the command:
   .predict [type] newvar [if] [in] [, single_options]

2. Observe the factor loading and discuss the rational of the final weight and whether they conflict with more some normative criteria.

3. Compute quintile groups based on the component score and assess the correlations with nutrition and education. Discuss if the extent to which the index is a good ad-hoc solution to measure socioeconomic status, the advantage and disadvantages of such a solution.

   Commands:
   . xtile newvar =   var, nq(5)
   . tabstat varlist [if] [in] [weight] [, options]

Exercise 2: Use EFA to inform your choices when constructing a MPI

In this second exercise we will explore how to use EFA to inform our choices on number of indicators, cluster by dimensions and weights. Remember that the analysis will be theoretically driven and we will not disregard the normative choices. Instead, we will assess the correlation and redundancy between indicators, the existence of latent variables and ways to group the indicators better for policy purposes. Consider following steps as a guide during your analysis:

1. Explore the correlation between the set of indicators that you have selected (assess different indicators with different deprivation cut-off). Can you identify high correlation (redundancy) among some indicators? Is there any clear pattern in the matrix of correlation? (Consider computing tetrachoric correlations).

2. Run a series of exploratory analysis to try identifying underlying variables. (Please see the command guide at the bottom of this document). Can you identify meaningful dimensions that could inform policy or be relevant for the purposes of your measure? Can you identify redundant variables that could be problematic?

3. If normatively or based on theory you determine that some variables are better clustered together in a dimension or should have a higher weight, try redefining the EFA without these indicators to assess the remaining indicators.

4. Discuss among your group and take a reasoned decision

Exercise 3 (optional): Validation of subjective scales with EFA and CFA

In this third exercise we will illustrate how factor analysis can be used during the process of validation of subjective scales. For this we will evaluate four scales from the module of psychological and subjective wellbeing from OPHI's missing dimensions, in particular: meaning of live, autonomy, social relatedness, and competence.

Data: 'OPHI2009_V3_ejercicio2.dta'
Variables and scales to validate: Meaning of live (mv3_a mv3_b mv3_c), Autonomy (mv4_a mv4_b mv4_c), Social Relatedness (mv5_a mv5_b mv5_c), Competence (mv6_a mv6_b mv6_c).
Additional variables: Happiness (*mv1),* Satisfaction (mv2_a)

Sample weight: factor

1. Explore the original variables with descriptive statistics.

2. Run a Exploratory Factor Analysis with the 12 variables and discuss: Do the scales measure the theoretical construct?

3. Generate the Alpha Cronbach coefficient to assess the internal consistency of the scales. Discussion in groups: Do the scales have internal consistency? Is it possible to improve the consistency by removing any item from the scale?

4. Run a Confirmatory Factor Analysis to evaluate the theoretical constructs (download and use the command confa). Discussion in groups: Are the theoretical construct confirmed?

Basic commands in STATA to implement Exploratory Factor Analysis

```
.factor varlist [if] [in] [weight] [, method options ]
     Method:         pf        principal factor; the default
                     pcf  principal-component factor
                     ipf       iterated principal factor
                     ml   maximum-likelihood factor

     Useful options: factors(#)   maximum number of factors to be
                     retained (usually, the analysis is run for the
                     first time without setting a number of factors)
                     blanks(#) display loadings as blanks when
                     |loadings| < # (this is useful but it is also
                     relevant that the analyst observes all loadings
                     in order to identify variable that might be
                     contributing with more than one factor
                     simultaneously).

. screeplot [eigvals] [, options ]

. findit fapara (in case the ADO file was not installed yet)
. fapara, reps(10)

.  rotate [, options]
     Useful options:  varimax  executes the method “varimax” during
                     the rotation; this is the most common orthogonal
                     method
                     promax executes the method “promax” during the
                     rotation; this is the most common oblique method
                     blanks(#)  display loadings as blank when
                     |loadings| < #; default is blanks(0)

 .predict [type] newvar [if] [in] [, single_options]
```

If the variables are dummies, the analysis can be implemented with the tetrachoric correlations matrix, by using the following commands:

```
. tetrachoric varlist [if] [in] [weight] [, options]
. matrix r = r(R)
.factormat r, n(#) {where # is the number of cases considered
```

Similarly, polychoric correlations matrix might be used if the variables are ordinal with few categories, to assume cardinality (it is also useful for mixed variables – dummies, ordinal and continuous variables). If the number of the categories of one of the variables is greater than 10, polychoric treats it is continuous, so the correlation of two variables that have 10 categories each would be simply the usual Pearson moment correlation found through the command correlate

```
.findit polychoric
. polychoric varlist [if] [in] [weight] [, options]
. matrix r = r(R)
.factormat r, n(#) {where # is the number of cases considered
```

It is worth noticing that the estimation of polychoric correlations is computationally demanding and that, under certain circumstances (sufficient categories for some type of analysis), the results are equivalent to those of the standard method, as mentioned previously.