Factor Analysis

Dr. Pankaj Kumar Singh
Assistant Professor, Department of Political Science, MGCUB

Introduction

- Factor analysis is mostly used for data reduction purposes
  - To get a small set of variables (preferably uncorrelated) from a large set of variables (most of which are correlated to each other)
  - To create indexes with variables that measure similar things (conceptually)

Cont...

- Frailty variables
  - Speed of walk
  - Speed of usual walk
  - Time to do chair stands
  - Arm circumference
  - Body mass index
  - Tricep skinfold thickness
  - Shoulder rotation
  - Upper extremity strength
  - Pinch strength
  - Grip strength
  - Knee extension
  - Hip extension
  - Time to do pegboard test

Cont...

- Other examples
  - Personality
  - Depression
  - Customer satisfaction
  - Woman’s autonomy

Applications of factor analysis

- Identification of underlying factors
  - Groups variables into homogeneous sets
  - Creates new variables
- Screening of variables
  - Identifies groupings to allow us to select one variable to represent many
  - Useful in regression (helps avoid collinearity)
- Summary
  - Allows us to describe many variables using few factors
- Clustering of objects
  - Helps us to put objects (people) into categories depending on their factor scores

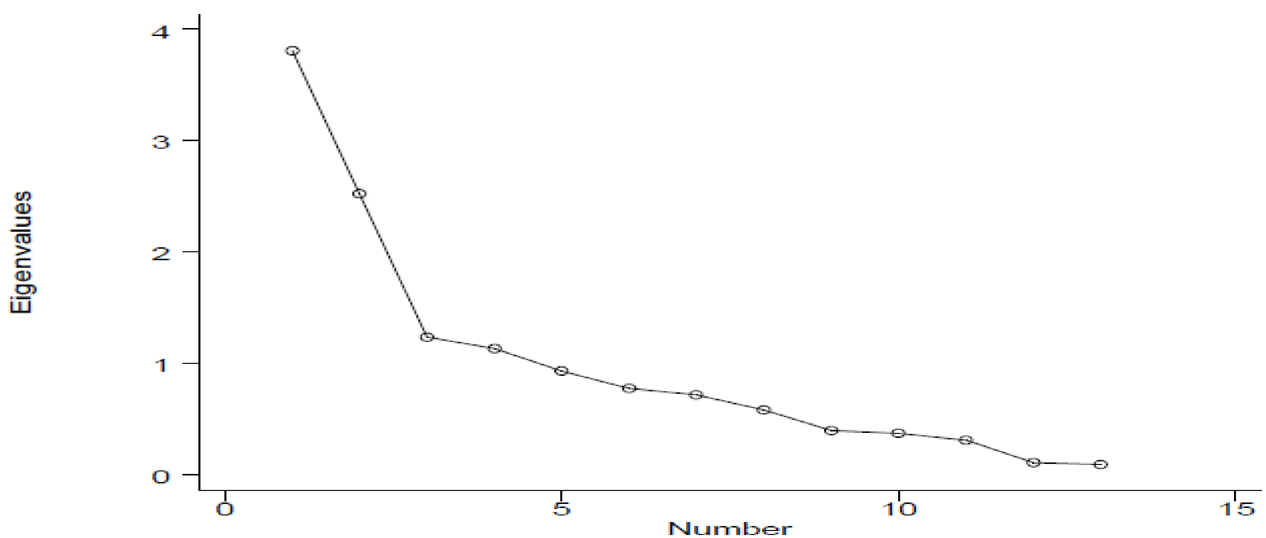
Some concepts

- **Eigenvalue**: is a measure of how much of the variance of the observed variables a factor explains
- So, if the factor for socioeconomic status (consisting of three variables i.e. income, education and occupation) had an eigenvalue of 2.3 it would explain as much variance as 2.3 of three variables
- The factors that explain the least amount of variance are generally discarded

Cont...

- **<u>Factor loading</u>**: The relationship of each variable to the underlying factor
- **<u>Scree plot</u>**: displays the eigenvalues associated with a factor in descending order versus the number of factor

Scree Plot for Frailty Example



Cont...

- Rotation: In principle components, the first factor describes most of variability
- After choosing number of factors to retain, we want to spread variability more evenly among factors
- To do this we 'rotate' factors:
  - Redefine factors such that loadings on various factors tend to be very high or very low
  - It makes sharper distinctions in the meanings of factors

Frailty example

Frailty Example

	Factors			
Variable	size	speed	Hand strength	Leg strength
arm_circ	**0.97**	-0.01	0.16	0.01
skinfld	**0.71**	0.10	0.09	0.26
fastwalk	-0.01	**0.94**	0.08	0.12
gripstr	0.19	0.10	**0.93**	0.10
pinchstr	0.26	0.09	**0.57**	0.19
upextstr	0.08	0.25	0.27	0.14
kneeext	0.13	0.26	0.16	**0.72**
hipext	0.09	0.09	0.14	**0.68**
shldrrot	0.01	0.22	0.14	0.26
pegbrd	-0.07	-0.33	-0.22	-0.06
bmi	**0.89**	-0.09	0.09	0.04
uslwalk	-0.03	**0.92**	0.07	0.07
chrstand	0.02	-0.43	-0.07	-0.18

Socioeconomic status example

Variables	Factor 1	Factor 2
Income	0.65	0.11
Education	0.59	0.25
Occupation	0.48	0.19
House value	0.38	0.60
Number of public parks in neighbourhood	0.13	0.57
Number of violent crimes per year in neighbourhood	0.23	0.55

Cont...

- The variable with the strongest association to the underlying latent variable factor 1 is income, with a factor loading of 0.65
- The other variables associated with factor 1 are education and occupation
- We could call factor 1 “individual socioeconomic status”
- We may call factor 2 “neighbourhood socioeconomic status”

Running factor analysis in STATA

- The command to run factor analysis in STATA
  - factor ideol equality owner respon competition, pcf
- After running factor you need to rotate the factor loads
  - rotate
- To create the new variables, after *factor, rotate* type predict
  - Predict factor1 factor2

Variables

Principal-components factoring

Since the sum of eigenvalues = total number of variables. Proportion indicate the relative weight of each factor in the total variance. For example, 1.54525/5=0.3090. The first factor explains 30.9% of the total variance

Total variance accounted by each factor. The sum of all eigenvalues = total number of variables.

When negative, the sum of eigenvalues = total number of factors (variables) with positive eigenvalues.

Kaiser criterion suggests to retain those factors with eigenvalues equal or higher than 1.

```
. factor  ideol equality owner respon competition, pcf
(obs=1125)

Factor analysis/correlation                      Number of obs    =     1125
    Method: principal-component factors          Retained factors =        2
    Rotation: (unrotated)                        Number of params =        9

    --------------------------------------------------------------------------
         Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      1.54524      0.21290            0.3090       0.3090
        Factor2  |      1.33235      0.49085            0.2665       0.5755
        Factor3  |      0.84149      0.12808            0.1683       0.7438
        Factor4  |      0.71341      0.14590            0.1427       0.8865
        Factor5  |      0.56751            .            0.1135       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(10) =  398.10 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

    ---------------------------------------------------
        Variable |  Factor1   Factor2 |   Uniqueness
    -------------+--------------------+----------------
           ideol |   0.4719    0.4019 |      0.6157
        equality |   0.4066    0.6424 |      0.4220
           owner |   0.6179   -0.5762 |      0.2861
          respon |   0.5807    0.4130 |      0.4922
     competition |   0.6619   -0.5056 |      0.3063
    ---------------------------------------------------
```

Cumulative shows the amount of variance explained by n+(n-1) factors. For example, factor 1 and factor 2 account for 57.55% of the total variance.

Difference between one eigenvalue and the next.

Uniqueness is the variance that is 'unique' to the variable and not shared with other variables. It is equal to 1 – communality (variance that is shared with other variables). For example, 61.57% of the variance in 'ideol' is not share with other variables in the overall factor model. On the contrary 'owner' has low variance not accounted by other variables (28.61%). Notice that the greater 'uniqueness' the lower the relevance of the variable in the factor model.

Factor loadings are the weights and correlations between each variable and the factor. The higher the load the more relevant in defining the factor's dimensionality. A negative value indicates an inverse impact on the factor. Here, two factors are retained because both have eigenvalues over 1. It seems that 'owner' and 'competition' define factor1, and 'equality', 'respon' and 'ideol' define factor2.

By default the rotation is varimax which produces orthogonal factors. This means that factors are not correlated to each other. This setting is recommended when you want to identify variables to create indexes or new variables without inter-correlated components

Same description as in the previous slide with new composition between the two factors. Still both factors explain 57.55% of the total variance observed.

The pattern matrix here offers a clearer picture of the relevance of each variable in the factor. Factor1 is mostly defined by 'owner' and 'competition' and factor2 by 'equality', 'respon' and 'ideol' .

This is a conversion matrix to estimate the rotated factor loadings (RFL):

RFL = Factor loadings * Factor rotation

```
. rotate

Factor analysis/correlation                    Number of obs    =      1125
    Method: principal-component factors        Retained factors =         2
    Rotation: orthogonal varimax (Kaiser off)  Number of params =         9

    --------------------------------------------------------------------------
         Factor  |     Variance   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      1.45169      0.02579            0.2903       0.2903
        Factor2  |      1.42590            .            0.2852       0.5755
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(10) =  398.10 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

    -------------------------------------------------
        Variable |  Factor1   Factor2 |   Uniqueness
    -------------+--------------------+--------------
           ideol |   0.0869    0.6138 |      0.6157
        equality |  -0.1214    0.7505 |      0.4220
           owner |   0.8446   -0.0218 |      0.2861
          respon |   0.1610    0.6941 |      0.4922
     competition |   0.8307    0.0603 |      0.3063
    -------------------------------------------------

Factor rotation matrix

    --------------------------------
                 |  Factor1  Factor2
    -------------+------------------
         Factor1 |   0.7487   0.6629
         Factor2 |  -0.6629   0.7487
    --------------------------------
```

```
predict factor1 factor2 /*or whatever name you prefer to identify the factors*/
```

```
. predict factor1 factor2
(regression scoring assumed)

Scoring coefficients (method = regression; based on varimax rotated factors)
```

Variable	Factor1	Factor2
ideol	0.02868	0.42832
equality	-0.12258	0.53541
owner	0.58610	-0.05873
respon	0.07591	0.48119
competition	0.57225	-0.00014

These are the regression coefficients used to estimate the individual scores (per case/row)

Variables

Name	Label
e033	self positioning in political scale
e035	Income equality
e036	private vs state ownership of bus...
e037	government responsibility
e039	competition good or harmful
ideol	Self positioning in political scale
equality	Income equality
owner	State vs private ownership of bus...
respon	Government vs individual respons...
competition	Competition harmful or good
f1	Scores for factor 1
f2	Scores for factor 2
f1a	Scores for factor 1
f2a	Scores for factor 2
factor1	Scores for factor 1
factor2	Scores for factor 2

Cont...

- Another option could be to create indexes out of each cluster or variables.
- For example, ‘owner’ and ‘competition’ define one factor. These two can be aggregated to create a new variable to measure ‘market oriented attitudes’
- ‘ideol’, ‘equality’ and ‘respon’ can be aggregated to create a new variable to measure ‘egalitarian attitudes’
- The two new variables can be created as
  – gen market = (owner + competition)/2
  – gen egalitarian = (ideol + equality + respon)/3

Choosing number of factors

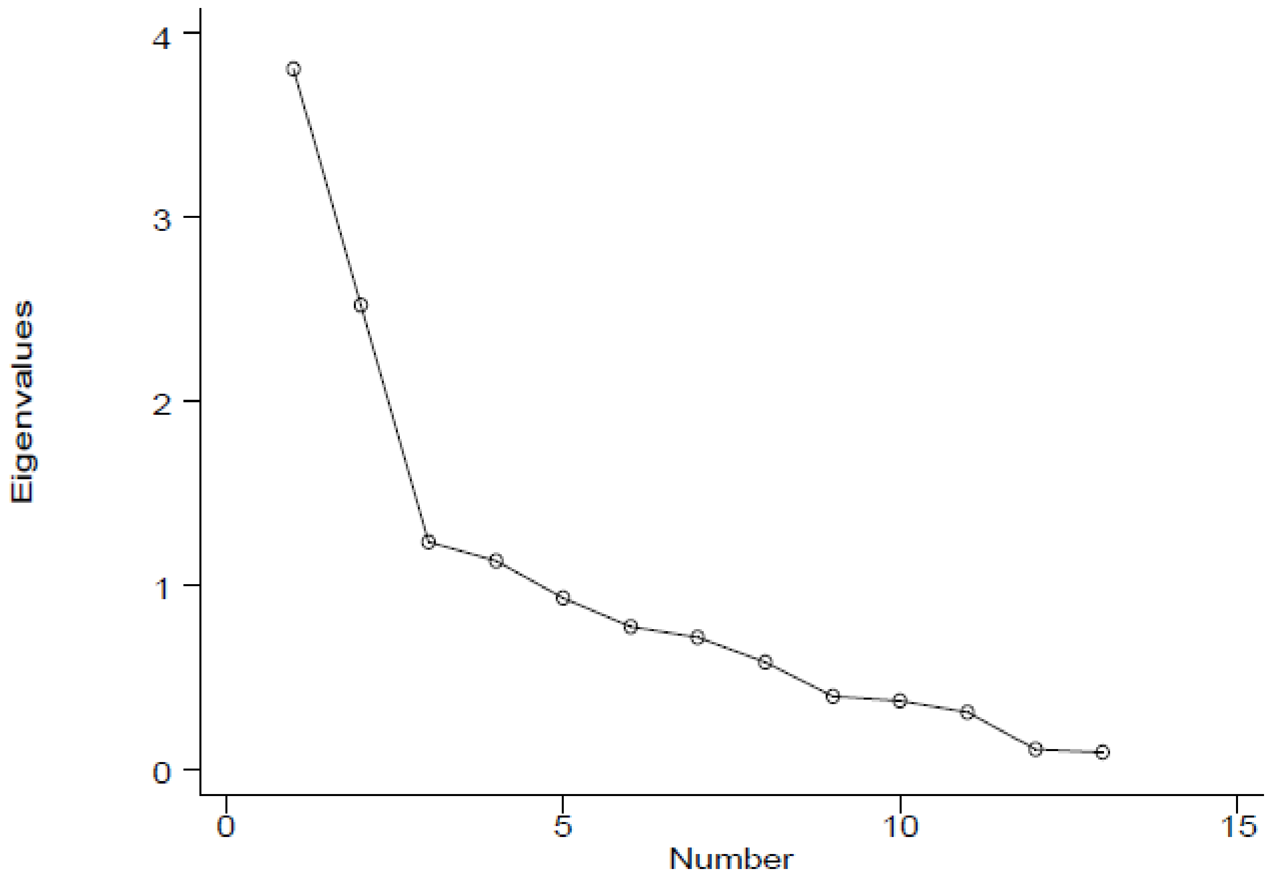
- To select how many factors to use, consider eigenvalues from a principal components analysis
- Rules to go by:
  - Number of eigenvalues > 1
  - Scree plot
  - % variance explained
  - comprehensibility

Frailty Example

(principal components; 13 components retained)

Component	Eigenvalue	Difference	Proportion	Cumulative
1	3.80792	1.28489	0.2929	0.2929
2	2.52303	1.28633	0.1941	0.4870
3	1.23669	0.10300	0.0951	0.5821
4	1.13370	0.19964	0.0872	0.6693
5	0.93406	0.15572	0.0719	0.7412
6	0.77834	0.05959	0.0599	0.8011
7	0.71875	0.13765	0.0553	0.8563
8	0.58110	0.18244	0.0447	0.9010
9	0.39866	0.02716	0.0307	0.9317
10	0.37149	0.06131	0.0286	0.9603
11	0.31018	0.19962	0.0239	0.9841
12	0.11056	0.01504	0.0085	0.9927
13	0.09552	.	0.0073	1.0000

Scree Plot for Frailty Example



5 Factors, Unrotated

Factor Loadings

Variable	1	2	3	4	5
arm_circ	**0.59934**	**0.67427**	-0.26580	-0.04146	0.02383
skinfld	**0.62122**	**0.41768**	-0.13568	0.16493	0.01069
fastwalk	**0.57983**	**-0.64697**	**-0.30834**	-0.00134	-0.05584
gripstr	**0.57362**	0.08508	**0.31497**	**-0.33229**	-0.13918
pinchstr	**0.55884**	0.13477	**0.30612**	-0.25698	-0.15520
upextstr	**0.41860**	-0.15413	0.14411	-0.17610	0.26851
kneeext	**0.56905**	-0.14977	0.26877	**0.36304**	-0.01108
hipext	**0.44167**	-0.04549	**0.31590**	**0.37823**	-0.07072
shldrrot	**0.34102**	-0.17981	0.19285	-0.02008	**0.31486**
pegbrd	**-0.37068**	0.19063	0.04339	0.12546	-0.03857
bmi	**0.51172**	**0.70802**	-0.24579	0.03593	0.04290
uslwalk	**0.53682**	**-0.65795**	**-0.33565**	-0.03688	-0.05196
chrstand	**-0.35387**	**0.33874**	0.07315	-0.03452	0.03548

5 Factors, Rotated

(varimax rotation)

Rotated Factor Loadings

Variable	1	2	3	4	5
arm_circ	-0.00702	**0.93063**	0.14300	0.00212	0.01487
skinfld	0.11289	**0.71998**	0.09319	0.25655	0.02183
fastwalk	**0.91214**	-0.01357	0.07068	0.11794	0.04312
gripstr	0.13683	0.24745	**0.67895**	0.13331	0.08110
pinchstr	0.09672	0.28091	**0.62678**	0.17672	0.04419
upextstr	0.25803	0.08340	0.28257	0.10024	**0.39928**
kneeext	0.27842	0.13825	0.16664	**0.64575**	0.09499
hipext	0.11823	0.11857	0.15140	**0.62756**	0.01438
shldrrot	0.20012	0.01241	0.16392	0.21342	**0.41562**
pegbrd	**-0.35849**	-0.09024	-0.19444	-0.03842	-0.13004
bmi	-0.09260	**0.90163**	0.06343	0.03358	0.00567
uslwalk	**0.90977**	-0.03758	0.05757	0.06106	0.04081
chrstand	**-0.46335**	0.01015	-0.08856	-0.15399	-0.03762