# Getting Started in Factor Analysis

(v. 1.5)

***Oscar Torres-Reyna***

*otorres@princeton.edu*

August 2008

http://www.princeton.edu/~otorres/

Factor analysis is used mostly 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).

*Two types of factor analysis*

*Exploratory*

It is exploratory when you do not have a pre-defined idea of the structure or how many dimensions are in a set of variables.

*Confirmatory*.

It is confirmatory when you want to test specific hypothesis about the structure or the number of dimensions underlying a set of variables (i.e. in your data you may think there are two dimensions and you want to verify that).

# Factor analysis: step 1

To run factor analysis use the command `factor` (type `help factor` for more details).

Variables

Principal-components factoring

```
. 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
-------------+--------------------+-----------------------
```

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.

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

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.

# Factor analysis: step 2 (final solution)

After running `factor` you need to rotate the factor loads to get a clearer pattern, just type `rotate` to get a final solution.

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
```

**NOTE**: If you want the factors to be correlated (oblique rotation) you need to use the option `promax` after `rotate`:
`rotate, promax`
Type `help rotate` for details. See http://www.ats.ucla.edu/stat/stata/output/fa_output.htm for more info.

Thank you to Jeannie-Marie S. Leoutsakos for useful feedback.

# Factor analysis: step 3 (predict)

To create the new variables, after `factor`, `rotate` you type `predict`.

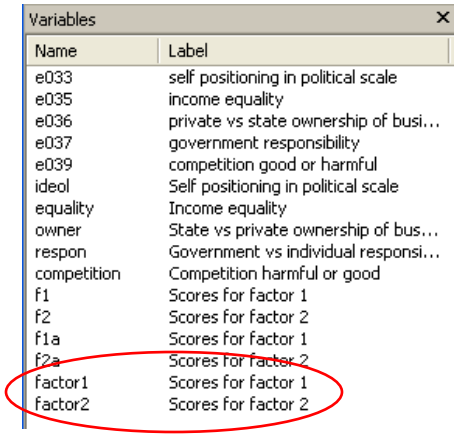
```
predict factor1 factor2 /*or whatever name you prefer to identify the factors*/
```

```
. 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)




Another option (called *naïve* by some) could be to create indexes out of each cluster of variables. For example, 'owner' and 'competition' define one factor. You could aggregate these two to create a new variable to measure 'market oriented attitudes'. On the other hand you could aggregate 'ideol', 'equality' and 'respon' to create an index to measure 'egalitarian attitudes'. Since all variables are in the same valence (liberal for small values, capitalist for larger values), we can create the two new variables as

```
gen market = (owner + competition)/2
gen egalitatiran = (ideol + equality + respon)/3
```

# Factor analysis: sources/references

The main sources/references for this section are:

**Books**

- *Factor Analysis in International Relations. Interpretation, Problem Areas and Application* / Vincent, Jack. University of Florida Press, Gainsville, 1971.
- *Factor Analysis. Statistical Methods and Practical Issues* / Kim Jae-on, Charles W. Mueller, Sage publications, 1978.
- *Introduction to Factor Analysis. What it is and How To Do It* / Kim Jae-on, Charles W. Mueller, Sage publications, 1978.
- *Statistics with STATA (updated for version 9)* / Hamilton, Lawrence C. Thomson Books/Cole, 2006.

**Online**

- StatNotes: http://faculty.chass.ncsu.edu/garson/PA765/factor.htm
- StatSoft: http://www.statsoft.com/textbook/stfacan.html
- UCLA Resources: http://www.ats.ucla.edu/stat/stata/output/fa_output.htm