Vanderbilt University
Leadership, Policy and Organizations
Class Number 9953
Spring 2021

Factor Analysis

Factor analysis is utilized when the analyst suspects that there is an unobserved but still important set of characteristics of individuals that lead them to act in certain ways or have certain attitudes. There are two types of factor analysis, which differ in philosophy but not in the actual application of methods:

Also, there are two broad approaches to this technique: one seeks to estimate components, while the other seeks to estimate factors. The difference is that techniques to recover components analyze all of the variance in the underlying items, while techniques to recover factors analyze only shared variance, while varience that is due to error or specific to the underlying items is (supposed to be) eliminated. We're going to focus on factor analysis, as that is usually what the analyst wants to do. In application, the two approaches yield similar but not identical answers.

Confirmatory Factor Analysis is used when the analyst has a strong theoretical construct that the expect to see "show up" in the data.

Exploratory Factor Analysis is used when the analyst has not *a priori* expectations about the factors that may or may not be in the data, but instead seeks to understand what they might be.

For the purposes of these notes, we'll focus on the latter.

The basic model for factor analysis posits that an individual i's response to a survey item j, $x_{ij}, (i = 1 \ldots n; j = 1 \ldots k)$ can be thought of as being driven by a set of p factors, each of which contributes in part to that response.

$$x_{ij} = \lambda_{j1}\xi_{i1} + \lambda_{j2}\xi_{i2} + \ldots + \lambda_{jp}\xi_{ip} + \delta_{ij}$$

Where λ are the factor loadings on each variable, and ξ are the various unobserved factors.

When we observe a correlation matrix, the theory behind factor analysis states that we should expect that the correlations observed are driven by a set of latent factors, as above. Factor analysis seeks to extract those factors from the correlation matrix in such a way that the factors are independent of one another, but can reproduce the correlation matrix.

We'll go over three models to estimate the factors that contribute to two different sets of questions, one of which asks members of the public what they consider to be the most important things that colleges can teach students, and another which asks people what they think college administrators should work on.

Principal Factors

The method of common factors seeks to find the smallest number of factors which can account for the covariance in a set of variables. This is the most computationally "cheap" and oldest method, but it is limited. In particular, principal factors focuses only on the commonalities among variables, ignoring any unit-level variance. Another way of thinking about this is that principal factors is only really trying to explain what's going on in the variance-covariance between variables, not with trying to explain the variance in the data itself.

To run a principal factors analysis in Stata:

```
. factor `students´, ipf factor(3)
(obs=949)

Factor analysis/correlation                        Number of obs    =      949
    Method: iterated principal factors             Retained factors =        3
    Rotation: (unrotated)                          Number of params =       21

    --------------------------------------------------------------------------
         Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      1.65621      1.28327            0.7610       0.7610
        Factor2  |      0.37294      0.22571            0.1714       0.9324
        Factor3  |      0.14724      0.09293            0.0677       1.0000
        Factor4  |      0.05431      0.04166            0.0250       1.0250
        Factor5  |      0.01265      0.01427            0.0058       1.0308
        Factor6  |     -0.00162      0.01586           -0.0007       1.0301
        Factor7  |     -0.01748      0.03045           -0.0080       1.0220
        Factor8  |     -0.04793            .           -0.0220       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(28) =  699.30 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

    -----------------------------------------------------------
        Variable |  Factor1   Factor2   Factor3 |   Uniqueness
    -------------+------------------------------+--------------
             q35 |   0.4975    0.2074   -0.0173 |      0.7092
             q36 |   0.3944    0.0846    0.2372 |      0.7810
             q37 |   0.4015    0.1525   -0.1199 |      0.8011
             q38 |   0.4817    0.1869   -0.0111 |      0.7329
             q39 |   0.5137   -0.4362   -0.1472 |      0.5241
             q40 |   0.5352   -0.0799    0.0057 |      0.7071
             q41 |   0.4117   -0.2013    0.2037 |      0.7485
             q42 |   0.3743    0.1655   -0.1139 |      0.8195
    -----------------------------------------------------------
```

The first table in the results reports the various factors estimated, their eigenvalues, and the proportion of variance in the items associated with that factor, both for that factor and cumulatively. In our example, there's only really one decent factor. A common rule is to only keep factors with eigenvalues greater than 1.

Stata next reports the factor loadings on each of the items. As we can see, the factor loadings are only marginally high for the first factor, and decrease from there. There is a strong negative relationship between factor 1 and factor 2 on question 39.

Principal Components

The method of principal components overcomes the problems association with principal factors by seeking the set of factors that are primarily correlated with the set of response, and importantly, uncorrelated with one another. What principal components does that principal factors does not is attempt to maximize the variance at the unit level. From Tabachnick and Fidell “the principal component is the linear combination of observed variables that maximally separates subjects by maximizing the variance of their component scores” (p. 664).

```
.
. factor `students´, pcf factor(3)
(obs=949)

Factor analysis/correlation                        Number of obs    =      949
    Method: principal-component factors           Retained factors =        2
    Rotation: (unrotated)                          Number of params =       15

    --------------------------------------------------------------------------
         Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      2.37332      1.32899            0.2967       0.2967
        Factor2  |      1.04433      0.15760            0.1305       0.4272
        Factor3  |      0.88673      0.06601            0.1108       0.5380
        Factor4  |      0.82072      0.03128            0.1026       0.6406
        Factor5  |      0.78943      0.06295            0.0987       0.7393
        Factor6  |      0.72649      0.02910            0.0908       0.8301
        Factor7  |      0.69739      0.03579            0.0872       0.9173
        Factor8  |      0.66160            .            0.0827       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(28) =  699.30 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

    -------------------------------------------------
        Variable |  Factor1   Factor2 |   Uniqueness
    -------------+--------------------+--------------
             q35 |   0.5971    0.2762 |      0.5672
             q36 |   0.4945   -0.0220 |      0.7550
             q37 |   0.5099    0.3493 |      0.6180
             q38 |   0.5857    0.2566 |      0.5911
             q39 |   0.5378   -0.4898 |      0.4708
             q40 |   0.6332   -0.2249 |      0.5485
             q41 |   0.4970   -0.5601 |      0.4393
             q42 |   0.4816    0.4190 |      0.5925
    -------------------------------------------------
```

This is giving us similar results to the above, but with the rotation we have two factors whose eigenvalues exceed 1.

Maximum Likelihood

Maximum likelihood approaches to factor analysis seeks to estimate the parameters in the above equation, subject to a set of identifying constraints. The Maximum Likelihood method has the most promising theoretical properties for identifying the principal factors λ, but it also has a tendency to run into boundary conditions, known as a Heywood case.

```
.
```

```
.   factor `students´, ml factor(3)
(obs=949)
Iteration 0:   log likelihood = -29.559081
Iteration 1:   log likelihood = -2.5809681
Iteration 2:   log likelihood = -2.2833716
Iteration 3:   log likelihood = -2.2690829
Iteration 4:   log likelihood = -2.2681242
Iteration 5:   log likelihood = -2.2680494
Iteration 6:   log likelihood = -2.2680434

Factor analysis/correlation                        Number of obs    =      949
    Method: maximum likelihood                     Retained factors =        3
    Rotation: (unrotated)                          Number of params =       21
                                                   Schwarz´s BIC    =    148.5
    Log likelihood = -2.268043                     (Akaike´s) AIC   =  46.5361

    --------------------------------------------------------------------------
         Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      1.64836      1.26417            0.7553       0.7553
        Factor2  |      0.38419      0.23423            0.1760       0.9313
        Factor3  |      0.14996            .            0.0687       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(28) =  699.30 Prob>chi2 = 0.0000
    LR test:   3 factors vs. saturated:  chi2(7)  =    4.51 Prob>chi2 = 0.7195

Factor loadings (pattern matrix) and unique variances

    -----------------------------------------------------------
        Variable |  Factor1   Factor2   Factor3 |   Uniqueness
    -------------+------------------------------+--------------
             q35 |   0.4797    0.2484   -0.0338 |      0.7070
             q36 |   0.3797    0.1386    0.2284 |      0.7845
             q37 |   0.3876    0.1721   -0.1195 |      0.8059
             q38 |   0.4666    0.2282   -0.0295 |      0.7293
             q39 |   0.5536   -0.4084   -0.1010 |      0.5165
             q40 |   0.5397   -0.0309    0.0191 |      0.7074
             q41 |   0.4227   -0.1448    0.2376 |      0.7438
             q42 |   0.3601    0.1812   -0.1204 |      0.8230
    -----------------------------------------------------------
```

Post-estimation

To report correlations between the factors reported out in your analysis and the items, use the following command:

```
. estat structure

Structure matrix: correlations between variables and common factors

    --------------------------------------------
        Variable |  Factor1   Factor2   Factor3
    -------------+------------------------------
             q35 |   0.4797    0.2484   -0.0338
             q36 |   0.3797    0.1386    0.2284
             q37 |   0.3876    0.1721   -0.1195
             q38 |   0.4666    0.2282   -0.0295
             q39 |   0.5536   -0.4084   -0.1010
             q40 |   0.5397   -0.0309    0.0191
             q41 |   0.4227   -0.1448    0.2376
             q42 |   0.3601    0.1812   -0.1204
    --------------------------------------------
```

You can also create new variables based on the factors estimated in your model like so:

```
. predict studt_*, bartlett
```

The `studt` is a stub, indicating a prefix for all of the factors to be predicted. There are two methods for prediction, a regression based method and “Bartlett’s” method. Bartlett’s method is known to be unbiased, but can be inaccurate (more variable).

Graphics

Three kinds of graphs are helpful for understanding factor analysis: a factor loading plot, a score plot, and a scree plot.

Loading Plots plot each variable relative to each factor, showing which variables load most heavily on each factor. These are used to show which items are most closely related to each factor.

Score Plots give a scatterplot of the predicted score for each individual against one or more other factors. These are used to show how the factors relate to one another.

Scree Plots plot the eigenvalues of each factor as a function of the number of factors. These are used to show how well the various factors fit the data (higher eigenvalues being better).