FACTOR ANALYSIS

INTRODUCTION

Factor analysis is a statistical technique used widely in psychology and the social sciences for the purpose of identifying unobservable constructs (factors) from a set of observables. There are two primary uses of factor analysis:

1. Exploring how many different factors are captured by the data, which measures are capturing which factors, and whether we can reduce the number of measures we use and still capture the same constructs. This is exploratory factor analysis, or EFA, which I focus on in this post.

2. Performing checks that a set of measures captures the constructs the researcher intended to capture. This is done using confirmatory factor analysis, or CFA.

The capabilities of EFA are best explained through an example. Suppose you are a recruitment agency, performing tests on job applicants and providing information about applicant quality to employers. You capture all school qualification scores and conduct a long series of tests: a basic numeracy test, a basic verbal test, logic puzzles, an IQ test, etc. You provide all these results to potential employers. But employers are confused when presented with this much information. They don't want to compare ~20 numbers across 100 people. In addition, administering all these tests is costly and time consuming for applicants. They may choose a different recruitment agency. EFA will:

1) Show you how the different tests are related to each other

2) Give you a set of underlying qualities (factors) measured by the tests

3) Enable you to reduce the number of tests you do, without losing the ability to capture these underlying qualities

Having performed EFA, you were able to identify three underlying candidate characteristics: creativity, verbal ability and numeracy. This made your candidate reports for employers much more accessible. In addition, you were able to identify redundant variables; once you had applicants' school results, you only had to perform two additional tests to capture all three characteristics.

As this example illustrates, EFA can be a very useful tool. For researchers hoping to capture unobservable constructs, such as psychological or personality traits, EFA can help to legitimise the measures used and reduce costs by identifying redundant measures. So, how does it work?

EXPLANATION

It is very hard to establish what is driving correlations when presented with a correlation matrix of a large number of measures. EFA works by asking what might account for all of these correlations. In other words, EFA is a method for investigating whether a number of variables of interest $Y_1, Y_2, \ldots, Y_k$, are linearly related to a smaller number of unobservable factors $F_1, F_2, \ldots, F_k$. For example:

$$Y_1 = \beta_{10} + \beta_{11}F_1 + \beta_{12}F_2 + e_1$$

$$Y_2 = \beta_{20} + \beta_{21}F_1 + \beta_{22}F_2 + e_2$$

$$Y_3 = \beta_{30} + \beta_{31}F_1 + \beta_{32}F_2 + e_3$$

Coefficients β_{ij} are referred to as factor loadings, of variable Y_i on factor F_j. As factors are unobservable, we cannot simply estimate the loadings by regression. Instead, we use factor analysis models. The simplest example is one which satisfies the following assumptions:

1. The error terms e_i are independent of one another, and such that $E(e_i) = 0$ and $Var(e_i) = \sigma_i^2$
2. The unobservable factors F_j are independent of one another and of the error terms, and are such that $E(F_j) = 0$ and $Var(F_j) = 1$

Note, more complex models allow for Assumption 2 to be relaxed; I will come back to this. The assumptions imply that the Variance of Y_i consists of:

$$Var(Y_i) = \beta_{i1}^2 + \beta_{i2}^2 + \sigma_i^2$$

The $\beta_{i1}^2 + \beta_{i2}^2$ component is called communality, or the part that is explained by the common factors F_1 and F_2; σ_i^2 is the specific variance, or the part of the variance that is not accounted for by the common factors.

The final term to introduce is uniqueness, which is the percentage of the variance that is not explained by the common factors, or $1 - communality$. Under the assumptions of the model, uniqueness is σ_i^2. However, if the assumptions are violated and uniqueness is not pure measurement error, it is accounting for something particular to the variable. The greater the uniqueness, the more likely that it is more than just measurement error. Values more than 0.6 are usually considered high. If the uniqueness is high, then the variable is not well explained by the factors.

FACTOR ANALYSIS IN STATA:

The model can be solved using several methods in Stata, using the command -factor-. The syntax is:

factor *varlist* [*if*] [*in*] [*weight*] [, *method options*]

The command offers a choice of the following methods for estimating factor loadings based on the correlation matrix:

Method options:

- **pf** principal factor (the default). The factor loadings are computed using the squared multiple correlations as estimates of the communality
- **pcf** principal-component factor. Unlike pf, the communalities are assumed to be 1 - meaning there is no uniqueness. The model will not be appropriate if the proportion of the variation explained by unique variation is high
- **ipf** iterated principal factor. This method re-estimates the communalities iteratively
- **ml** maximum likelihood factor. This method assumes that the data are multivariate normal distributed

Example 1

Example 1 shows the output displayed when using the default method for estimating the factor structure of a set of measures designed to capture self-efficacy, or a person's judgement of how well they can deal with prospective situations. The log file shows output for this example using the different estimation methods.

The first column of the first table, Factor, will always equal the number of variables: in this case we tested 10 measures trying to capture self-efficacy. The eigenvalues in the second column show the total variance accounted for by each factor. The difference column shows the magnitude of the differences between one eigenvalue and the next. Proportion indicates the relative weight of each factor in the total variance: for example, Factor 1 accounts for: $2.72789/(2.72789+\ldots-0.16821) = 1.1351$of the variance. Cumulative column simply shows the cumulative proportion of the variance accounted for, always ending in 1. The second table provides the estimates of factor loadings and uniqueness.

There are several options within -factor- which are listed in the Stata manual; however, two options deserve particular mention at this stage. We can choose the number of factors from the first table, overriding the manual settings:

- **factors(#)** sets the maximum number of factors to be retained
- **mineigen(#)** sets the minimum value of eigenvalues to be retained. The default for all methods except pcf is a number very close to zero, meaning that factors associated with negative eigenvalues will not be retained. The default for pcf is 1 and this value is more common in literature. In Example 1, specifying mineigen(1) would lead to only Factor 1 being retained.

Example 1

```
. factor `efficacy'
(obs=11,061)

Factor analysis/correlation                      Number of obs    =     11,061
    Method: principal factors                    Retained factors =          4
    Rotation: (unrotated)                        Number of params =         34

    ------------------------------------------------------------------------------
        Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    ------------+-----------------------------------------------------------------
       Factor1  |      2.72789      2.38018            1.1351       1.1351
       Factor2  |      0.34771      0.32638            0.1447       1.2798
       Factor3  |      0.02132      0.00639            0.0089       1.2886
       Factor4  |      0.01493      0.08792            0.0062       1.2949
       Factor5  |     -0.07299      0.00729           -0.0304       1.2645
       Factor6  |     -0.08029      0.02486           -0.0334       1.2311
       Factor7  |     -0.10514      0.03006           -0.0438       1.1873
       Factor8  |     -0.13520      0.01156           -0.0563       1.1311
       Factor9  |     -0.14676      0.02145           -0.0611       1.0700
      Factor10  |     -0.16821            .           -0.0700       1.0000
    ------------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(45) = 2.0e+04 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

    ------------------------------------------------------------------------------
       Variable |  Factor1   Factor2   Factor3   Factor4 |   Uniqueness
    ------------+-----------------------------------------+--------------
     efficacy_1 |   0.3900    0.2620    0.0297    0.0357 |      0.7771
     efficacy_2 |   0.4580    0.0465    0.0269    0.0883 |      0.7796
    efficacy_3a |   0.4690    0.2256    0.0161   -0.0506 |      0.7263
     efficacy_4 |   0.5696   -0.1451    0.0622    0.0005 |      0.6506
     efficacy_5 |   0.5791   -0.2443    0.0438   -0.0104 |      0.6029
     efficacy_6 |   0.4335    0.2233    0.0164   -0.0155 |      0.7617
     efficacy_7 |   0.5969    0.0433    0.0071   -0.0511 |      0.6392
     efficacy_8 |   0.5628    0.0668   -0.0699    0.0182 |      0.6736
     efficacy_9 |   0.5843   -0.0110   -0.0891    0.0035 |      0.6505
    efficacy_10 |   0.5328   -0.2983   -0.0228    0.0000 |      0.6266
    ------------------------------------------------------------------------------
```

ROTATION:

The second table of Example 1 provides a set of solutions; however, there is no unique set of solutions as there are infinite sets of loadings which fit the same model. We seek a set of loadings that fit the observations equally well as the output of -factor- but are easier to interpret. This is achieved through a rotation and the command for this is -rotate-:

rotate [, *options*]

There are infinite possible rotations; however, there are certain methods which help to specify the kind of solution we are seeking. We can adapt our assumptions about factors being independent of each other at this stage; in the options you can specify:

- **oblique** For an oblique rotation to be applied, instead of the default, which is the orthogonal rotation. The factors before rotation are usually orthogonal, whereas the oblique rotated factors can be correlated

Depending on the choice between an orthogonal and an oblique rotation, the following rotation methods can be specified (this is a subset of the most commonly used rotations; for full details, consult the -rotate- chapter of the Stata manual):

- **varimax** This is the default for an orthogonal rotation. It seeks the rotated loadings that maximize the variance of the squared loadings for each factor. The idea is to make some loadings as large as possible, and the rest as small as possible in absolute value
- **quartimax** This is an alternative to varimax; it maximizes the variance of the squared loadings within the variables and tends to produce factors with high loadings for all variables
- **oblimin(#)** This option is suitable for both orthogonal and oblique rotations. The default is **oblimin(0)**, with values above 0 not recommended for oblique rotations.

Example 2: Final output

Based on the output from Example 1, the researcher decides to model the data using a two factor structure, performing an oblique rotation due to the likelihood of factors being correlated. The full output is displayed in the log file. After rotating, a further useful command is -sortl- which sorts the variables in order of factor loadings, making it easier to see which variables load in each factor. The final set of factor loadings is shown below.

Based on the results, the researcher would conclude that items 4, 5, 9 and 10 capture one construct while items 1, 2, 3a, 6, 7 and 8 capture a different construct. To get a sense of what the constructs could be, the researcher would explore the original measures. Perhaps some of the measures were negatively-worded, while some were positively worded. Alternatively, the two sets could be capturing slightly different aspects of self-efficacy, for example believing in yourself and coping when faced with adversities.

Example 2

Rotated factor loadings (pattern matrix) and unique variances sorted

Variable	Factor1	Factor2	Uniqueness
efficacy_10	0.6550	-0.0721	0.6271
efficacy_5	0.6171	0.0175	0.6049
efficacy_4	0.4940	0.1314	0.6545
efficacy_9	0.3432	0.3001	0.6585
efficacy_1	-0.0906	0.5235	0.7793
efficacy_3a	-0.0027	0.5222	0.7291
efficacy_6	-0.0201	0.5005	0.7622
efficacy_8	0.2387	0.3819	0.6788
efficacy_7	0.2859	0.3720	0.6419
efficacy_2	0.2036	0.3013	0.7881

Next, the researcher might consider whether some items are redundant, based on the factor loadings. Generally, loadings below 0.3 are considered low. Therefore, the researcher might choose to conduct this analysis dropping each of the lowest scoring items, 2 and 9. If the results remain similar, the researcher might be able to drop these in future studies, saving some valuable survey time.

Finally, the high level of uniqueness within the variables should remain a concern for the researcher, as it indicates that factors do not account for a very large proportion of the variance. This may throw doubt on the researcher's existing metrics and encourage them to seek alternative measures.

REFERENCES AND FURTHER READING

1. Stata (2019). Multivariate Statistics Reference Manual. Stata Press.
    a. mvfactor available at: https://www.stata.com/manuals13/mvfactor.pdf and the references within
    b. mvrotate available at: https://www.stata.com/manuals13/mvrotate.pdf and the references within
2. Paul Kline (1994). An easy guide to factor analysis. Routledge. This is a well written, light introduction to the key concepts
3. Oscar Torres-Reyna. Getting Started in Factor Analysis (using Stata 10). A short set of useful slides available at: https://dss.princeton.edu/training/Factor.pdf
4. Hoelzle & Meyer (2012). Exploratory factor analysis: Basics and beyond. Handbook of Psychology, Second Edition, 2.
5. Peter Tryfos, chapter on factor analysis available at: http://www.yorku.ca/ptryfos/methods.htm

Marta Grabowska, Research Assistant in Development Economics, Blavatnik School of Government

03 December 2019