Factor Analysis

Factor analysis is used to uncover the latent structure of a set of variables. It reduces attribute space from a large no. of variables to a smaller no. of factors and as such is a non dependent procedure.

Factor analysis could be used for any of the following purpose-

1. **To reduce a large no. of variable to a smaller no. of factors for modeling purposes,** where the large number of variables precludes modeling all the measures, individually. As such factor analysis is integrated in structural equation modeling, helping create the latent variables modeled by SEM (structure equation model).

2. To select a subset of variables from a large set based on which original variable have the highest correlations with the principal component factors.

3. To create a set of factors to be treated as uncorrelated variable as one approach to handling multicollinearity regression.

Assumptions:

Factor analysis is a part of the **Multiple General Linear Hypothesis (MGLH)**, family of procedures and makes many of the same assumptions as multiple regressions. Linear relationship interval or near-interval data, untruncated data, proper specification (relevant variable included extraneous are excluded), lack of high multicollinearity and multivariable normality for purpose of significant testing.
Factor analysis generates a table on which the rows and the observed row indicator variables and the columns are the factor or latent variables which explain as much of the variable in those variables as possible. The cells in this table are factor loadings and the meaning of the factors must be induced from seeing which variables are most heavily loaded on which factors this inferential process can be fraught with difficulty as diverse researchers impute different tables.

Methods

There are several different types of factor analysis-
1. Principal component method
2. Principal axes method
3. Summation method
4. Centroid method

Assumptions of factor analysis model

1) Measurement error has constant variance and is on average zero, i.e.,
   $var(e_i) = \sigma_i^2$
   $E[e_i] = 0$
2) No association between the factor and measurement error, $\text{cov}(F, e_j) = 0$
3) No association between errors, $\text{cov}(e_j, e_k) = 0$

4) **Local (i.e., conditional independence):** Given factor, observed variables are independent of one another, $\text{cov}(X_j, X_k \mid F) = 0$

Steps in Exploratory Factor Analysis:

1. Collect data: choose relevant variables.
2. Extract initial factors (via principal component).
3. Choose number of factors to retain.
4. Choose estimation method, estimate model.
5. Rotate and interpret.
6. (a) Decide on changes need to be made (e.g. drop items include items)
   (b) Repeat (4), (5).
7. Construct scales and use on further analysis.

Principal component method:

In this method, factors are selected one at a time such that each factor best fits the data. The first fraction is created such that it represents the most highly correlated set of variables.
Each subsequent selected factor explains less variance than its predecessor.
This procedure is continued till all the factors are selected. All the factors selected explain the largest amount of residual variance in the entire set of standardized response scores.

Centroid Method:

1) Obtain the correlation matrix.
2) Obtain grand matrix sum, row sum, column sum.
3) Calculate $N = \frac{1}{\sqrt{grandtotal}}$
4) Multiply each column sum with *N*, which gives the first factor loading.
5) To find the second factor loading, find the cross product matrix of factor 1 by testing first factor loading horizontally and vertically and then multiplying corresponding rows and columns.
6) Find the first factor residual matrix given as,
   Residuals = total variations- explained variations
   $= r_{ij} - l_{ij}$
7) Reflection:
   Reflection means that each test vector retains its length but extends in opposite directions. The major purpose of reflection is to get a reflected cost matrix having the highest possible grand total. This step is taken due to the reason that some of the factors loading is with grand total. The method of reflection is by trial and error. This can be done by changing the signs of the variables from positive to negative and negative to positive column wise and row wise. The outcome we get is a reflective residual correlation matrix.

8) Repeat steps from (3) to (7).

Methods for finding number of factors to be extracted

1) Thumb Rule: All the interrelated factors must explain at least as much as variances as an average variable. Check, if a variable is under a factor then the percentage of variable explaining variance should be less than the percentage of factor explaining.

2) Eigen Value Index: When the eigen value of a factor is less than 1, it explains less variance then the variables included in the factor itself such a factor should not be considered.

3) Fruckter Formula:

$$\text{Number of factors} = \frac{(2n-1)-\sqrt{8n+1}}{2}$$

Where n is the number of variables included in the study.

4) Residual correlation matrix method: In this method, the residual correlation is observed and if it is soon that most of the correlation coefficient in this matrix are zero, and then the extraction of the factor can be determinate.

5) Scree Plot test: This method is to decide about the number of factors to be retained from the extracted factors. The test determines which of the extracted factors are actually contributing variance and does not measure random errors. The number of factors is plotted against the proportion of variance. It extracts in the order of the extracted factors.

Standardization of Responses:

$$\hat{X}_i = \frac{X_i - \overline{X}}{\sigma}$$

where X_i is a value corresponding to a response and σ is the variance.

Factor Loading:

It is the correlation between a factor and a variable. It helps interpret the meaning of a factor by indicating how well the factor fits the standardized responses to a variable. **The greater the value of factor loading the better is the fit of the factor to** the data from the concerned statement. All variables load on all factors but they load highly on some specific factors. Range is ± 1

There may be some variables which may by loading highly to more than two factors, decide in which factors which variables are to be kept.

Eigen Values: It is the measurement of the amount of variants explained by a factor. A factor eigen value is the sum of the square of its factor loading.

Communality: It indicates the proportion of variance in the responses to the statement which is explained by the identified factors.

Percentage of Variance:

$$= \frac{\text{eigen value of the factor}}{\text{sum of all eigen values}} \times 100$$

If any variable is not combined in any of the groups, then it can be left or can be considered as another factor. To remove variable which is totally different use rotation i.e., we are basically changing its direction.

Rotation of Factors: For the purpose of simplifying the interpretation of obtained factors and to increase the number of high and low positive loadings in the columns of a factor, factor rotation is used. There are two methods for this:

1) Orthogonal Rotation/ Variance Rotation: Here factors are rotated such that the original factors as well as rotated factors are orthogonal. The line between the factors axis remains $90^\bullet$

2) Promax Rotation: The factors are rotated such that the line between original and rotated factors is more than or less than $90^\bullet$

Advantages of Factor Analysis:

1. Both objective and subjective attributes can be used.
2. It can be used to identify the hidden dimensions or constraints which may or may not be apparent from direct analysis.
3. It is not extremely difficult to do and at the same time its inexpensive and gives accurate results.
4. There is flexibility in naming and using dimensions.

Disadvantages of Factor Analysis:

1. The usefulness depends on the researcher's ability to develop a complete and accurate set of product attributes. If important attributes are missed the value of procedure is reduced accordingly.
2. Naming of the factors can be difficult multiple attributes can be highly correlated with no apparent reasons.
3. If the observed variables are completely unrelated the factor analysis is unable to produce meaningful pattern.
4. It is not possible to know factors actually represents, only theory can help inform the researcher's on this.