Question #66589, Economics / Microeconomics

State the various uses of factor analysis for conducting research in social sciences.
Discuss the methods for estimating the parameters of a factor model.

Answer:

Factor analysis is most familiar to researchers as an *exploratory* tool for unearthing the basic empirical concepts in a field of investigation. Representing patterns of relationship between phenomena, these basic concepts may corroborate the reality of prevailing concepts or may be so new and strange as to defy immediate labeling. Factor analysis is often used to discover such concepts reflecting unsuspected influences at work in a domain. The delineation of these interrelated phenomena enables generalizations to be made and hypotheses posed about the underlying influences bringing about the relationships. For example, if a political scientist were to factor the attributes and votes of legislators and were to find a pattern involving urban constituencies and liberal votes, he could use this finding to develop a theory linking urbanism and liberalism. The ability to *relate* data in a meaningful fashion is a prime aspect of induction and, for this, factor analysis is useful and efficient.

There are a number of different methods that can be used for estimating factor scores from the data. These include:

- Ordinary Least Squares
- Weighted Least Squares
- Regression method

Ordinary Least Squares

By default, this is the method that SAS uses if you use the principal component method of analysis. Unfortunately, SAS is a little bit vague about what it is doing here. Usually SAS will give you plenty of detail about how results are derived, but on this one it seems to be very vague.

Basically, we have our model and we look at the difference between the *j*th variable on the *i*th subject and its value under the factor model. The **L**'s are factor loadings and the **f** are our unobserved common factors. The following is performed done subject by subject.

So here, we wish to find the vector of common factors for subject *i*, or $\hat{\mathbf{f}}_i$, by minimizing the sum of the squared residuals:

$$\sum_{j-1}^{p} \varepsilon_{ij}^2 = \sum_{j-1}^{p}(y_{ij} - \mu_j - l_{j1}f_1 - l_{j2}f_2 - \cdots - l_{jm}f_m)^2 = (\mathbf{Y}_i - \boldsymbol{\mu} - \mathbf{L}\mathbf{f}_i)'(\mathbf{Y}_i - \boldsymbol{\mu} - \mathbf{L}\mathbf{f}_i)$$

This is like a least squares regression, except in this case we already have estimates of the parameters (the factor loadings), but wish to estimate the explanatory common factors. In matrix notation the solution is expressed as:

$$\hat{\mathbf{f}}_i = (\mathbf{L}'\mathbf{L})^{-1}\mathbf{L}'(\mathbf{Y}_i - \mu)$$

In practice, we substitute in our estimated factor loadings into this expression as well as the sample mean for the data:

$$\hat{\mathbf{f}}_i = \left(\hat{\mathbf{L}}'\hat{\mathbf{L}}\right)^{-1}\hat{\mathbf{L}}'(\mathbf{Y}_i - \bar{\mathbf{y}})$$

Using the principal component method with the unrotated factor loadings, this yields:

$$\hat{\mathbf{f}}_i = \begin{pmatrix} \frac{1}{\sqrt{\hat{\lambda}_1}}\hat{\mathbf{e}}_1'((\mathbf{Y}_i - \bar{\mathbf{y}})) \\ \frac{1}{\sqrt{\hat{\lambda}_2}}\hat{\mathbf{e}}_2'((\mathbf{Y}_i - \bar{\mathbf{y}})) \\ \vdots \\ \frac{1}{\sqrt{\hat{\lambda}_m}}\hat{\mathbf{e}}_m'((\mathbf{Y}_i - \bar{\mathbf{y}})) \end{pmatrix}$$

$\mathbf{e}_1$ through $\mathbf{e}_m$ are our first *m* eigenvectors.

Weighted Least Squares (Bartlett)

This alternative is similar to the Ordinary Least Squares method. The only real difference is that we are going to divide by the specific variances when we are taking the squared residual as shown below. This is going to give more weight, in this estimation, to variables that have low specific variances. Variables that have low specific variances are those variables for which the factor model fits the data best. We posit that those variables that have low specific variances give us more information regarding the true values for the specific factors.

Therefore, for the factor model:

$$\mathbf{Y}_i = \mu + \mathbf{L}\mathbf{f}_i + \varepsilon_i$$

we want to find $\hat{\mathbf{f}}_i$ that minimizes

$$\sum_{j=1}^{p}\frac{\varepsilon_{ij}^2}{\Psi_j} = \sum_{j=1}^{p}\frac{(y_{ij} - \mu_j - l_{j1}f_1 - l_{j2}f_2 - \cdots - l_{jm}f_m)^2}{\Psi_j} = (\mathbf{Y}_i - \mu - \mathbf{L}\mathbf{f}_i)'\Psi^{-1}(\mathbf{Y}_i - \mu - \mathbf{L}\mathbf{f}_i)$$

The solution is can be given by this expression where Ψ is the diagonal matrix whose diagonal elements are equal to the specific variances:

$$\hat{\mathbf{f}}_i = (\mathbf{L}'\Psi^{-1}\mathbf{L})^{-1}\mathbf{L}'\Psi^{-1}(\mathbf{Y}_i - \mu)$$

and can be estimated by substituting in the following:

$$\hat{\mathbf{f}}_i = \left(\hat{\mathbf{L}}'\hat{\Psi}^{-1}\hat{\mathbf{L}}\right)^{-1}\hat{\mathbf{L}}'\hat{\Psi}^{-1}(\mathbf{Y}_i - \overline{\mathbf{y}})$$

Regression Method

This method is used when you are calculating maximum likelihood estimates of factor loadings. What it involves is looking at a vector that includes the observed data, supplemented by the vector of factor loadings for the *i*th subject.

Joint distribution of the data ***Y**$_i$* and the factor ***f**$_i$* is

$$\begin{pmatrix} \mathbf{Y}_i \\ \mathbf{f}_i \end{pmatrix} \sim N\left[\begin{pmatrix} \mu \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{L}\mathbf{L}' + \Psi & \mathbf{L} \\ \mathbf{L}' & \mathbf{I} \end{pmatrix}\right]$$

Using this we can calculate the conditional expectation of the common factor score ***f**$_i$* given the data ***Y**$_i$* as expressed here:

$$E(\mathbf{f}_i|\mathbf{Y}_i) = \mathbf{L}'(\mathbf{L}\mathbf{L}' + \Psi)^{-1}(\mathbf{Y}_i - \mu)$$

This suggests the estimator by substituting in the estimates for the **L** and Ψ:

$$\hat{\mathbf{f}}_i = \hat{\mathbf{L}}'\left(\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Psi}\right)^{-1}(\mathbf{Y}_i - \overline{\mathbf{y}})$$

There is a little bit of a fix that often takes place to reduce the effects of incorrect determination of the number of factors. This tends to give you results that are a bit more stable.

$$\tilde{\mathbf{f}}_i = \hat{\mathbf{L}}'\mathrm{S}^{-1}(\mathbf{Y}_i - \overline{\mathbf{y}})$$

Sources:

https://www.hawaii.edu/powerkills/UFA.HTM

http://sites.stat.psu.edu/~ajw13/stat505/fa06/17_factor/14_factor_estimate.html