Analiza factorială

Analiza factorială descrie fiecare variabilă cauzală ca o combinație liniară de factori comuni (variabile latente) plus un factor unic sau specific, astfel:

$$\begin{cases} X_1 = l_{11}F_1 + l_{12}F_2 + ... + l_{1q}F_q + e_1 \\ X_2 = l_{21}F_1 + l_{22}F_2 + ... + l_{2q}F_q + e_2 \\ ... \\ X_m = l_{m1}F_1 + l_{m2}F_2 + ... + l_{mq}F_q + e_m \end{cases} \quad (1)$$

unde q este numărul factorilor ($q < m$), X_j, $j = \overline{1,m}$, sunt variabilele cauzale (variabilele observate) centrate sau standardizate, F_k, $k = \overline{1,q}$, sunt factorii comuni, e_j, $j = \overline{1,m}$ sunt factorii specifici iar l_{jk}, $j = \overline{1,m}$, $k = \overline{1,q}$ sunt coeficienții factoriali (factor loadings).

Matriceal, legătura dintre variabilele observate și factori se poate scrie astfel:

$X = F \cdot L^T + e$,

unde X este tabelul de observații, L^T este transpusa matricei coeficienților factoriali, $e = \begin{bmatrix} e_{11} & e_{12} & ... & e_{1m} \\ e_{21} & e_{22} & ... & e_{2m} \\ ... & & ... & \\ e_{n1} & e_{n2} & ... & e_{nm} \end{bmatrix}$ este matricea factorilor specifici, iar $F = \begin{bmatrix} f_{11} & f_{12} & ... & f_{1q} \\ f_{21} & f_{22} & ... & f_{2q} \\ ... & & ... & \\ f_{n1} & f_{n2} & ... & f_{nq} \end{bmatrix}$ este matricea factorilor comuni așezați pe coloane.

Ipotezele modelului

Factorii comuni sunt standardizați, prin urmare au varianță 1 și medie 0. Sunt construiți după principiul separației informaționale, deci sunt necorelați doi câte doi:

$R(F_i, F_j)=0$, $Cov(F_i, F_j)=0$, $i = \overline{1,q}$, $j = \overline{1,q}$, $i \neq j$.

Factorii specifici sunt de medie 0. Varianța factorilor specifici se notează cu $\psi_j = \text{var}(e_j)$, $j = \overline{1,m}$ și se numește varianța specifică.

Corelațiile și covarianțele dintre factorii specifici sunt 0, $R(e_i, e_j) = 0$, $Cov(e_i, e_j) = 0$, $i = \overline{1,m}$, $j = \overline{1,m}$, $i \neq j$.

Intre factorii comuni și factorii specifici nu există suprapunere de informație, deci sunt absolut necorelați între ei:

$R(F_i, e_j) = 0$, $Cov(F_i, e_j) = 0$, $i = \overline{1,q}$, $j = \overline{1,m}$.

Aceste ipoteze sunt necesare pentru a estima în mod unic parametrii modelului.

Având în vedere aceste ipoteze, varianța unei variabile observate este:

$$Var(X_j) = Var(l_{j1} \cdot F_1 + l_{j2} \cdot F_2 + ... + l_{jq} \cdot F_q + e_j) = l_{j1}^2 \cdot Var(F_1) + l_{j2}^2 \cdot Var(F_2) + ... + l_{jq}^2 \cdot Var(F_q) + Var(e_j) = \sum_{k=1}^{q} l_{jk}^2 + \psi_j .$$

Suma $h_j = \sum_{k=1}^{q} l_{jk}^2$ se numește ***comunalitatea*** variabilei X_j. Se observă că dacă variabilele X_j sunt standardizate, comunalitățile au valori cel mult egale cu 1. Comunalitatea reprezintă variabilitatea comună, datorată factorilor comuni.

Covarianța dintre două variabile observate X_i și X_j este:

$$Cov(X_i, X_j) = Cov(l_{i1}F_1 + ... + l_{iq}F_q + e_i, l_{j1}F_1 + ... + l_{jq}F_q + e_j) = \sum_{k=1}^{q} l_{ik} l_{jk} .$$

Covarianța dintre o variabilă observată și un factor comun este:

$Cov(X_i, F_j) = Cov(l_{i1}F_1 + ... + l_{iq}F_q + e_i, F_j) = l_{ij}$.

Ținând cont de aceste relații, matricea de covarianță a tabelului de observații se poate scrie:

$V = L \cdot L^T + \psi$,

unde $\psi = \begin{bmatrix} \psi_1 & 0 & ... & 0 \\ 0 & \psi_2 & ... & 0 \\ ... & & ... & \\ 0 & 0 & ... & \psi_m \end{bmatrix}$ este matricea diagonală a varianțelor factorilor specifici.

Estimarea existenței factorilor

Ipoteza existenței unor factori latenți este indusă de o corelație semnificativă dintre variabilele cauzale. Variabilitatea comună a variabilelor cauzale conduce la ideea explicării acesteia prin existența unor factori ascunși. Deci pe baza matricei de corelații a variabilelor cauzale se poate estima existența factorilor comuni.

Există mai multe teste prin care se poate testa ipoteza existenței factorilor pe baza matricei de corelații.

Testul de sfericitate Bartlett

Este un test χ^2 care compară matricea de corelații cu matricea unitate. Ipotezele testului:

H0: Nu există factori

H1; Există cel puțin un factor comun

În cazul unor variabile absolut necorelate, determinantul matricei de corelații este 1.

Statistica testului:

$$\chi^2 = -\left(n-1-\frac{2\cdot m+5}{6}\right)\ln|R|,$$

unde R este matricea de corelații, n numărul de instanțe și m numărul de variabile. Aceste valori urmează o distribuție χ^2 cu $m(m\text{-}1)/2$ grade de libertate.

Dacă $\chi^2 > \chi_C^2(\alpha; m\cdot(m-1)/2)$,

ipoteza nulă este respinsă cu un nivel de încredere 1-α.

Indicele KMO (Kaiser-Meyer-Olkin)

Se bazează tot pe matricea de corelații.

Indicele KMO global se calculează astfel:

$$KMO = \frac{\sum_i \sum_{j\neq i} r_{ij}^2}{\sum_i \sum_{j\neq i} r_{ij}^2 + \sum_i \sum_{j\neq i} a_{ij}^2},$$

unde r_{ij} este coeficientul de corelație liniară dintre variabilele X_i și X_j iar a_{ij} reprezintă coeficientul de corelație parțială dintre X_i și X_j.

Indicii *KMO* pentru fiecare variabilă se calculează astfel:

$$KMO_j = \frac{\sum_{i\neq j} r_{ij}^2}{\sum_{i\neq j} r_{ij}^2 + \sum_{i\neq j} a_{ij}^2},\ j = \overline{1,m}.$$

Acești indici arată care variabile sunt mai puțin corelate cu celelalte, deci oferă mai puțină variabilitate comună.

Coeficientul de corelație parțială dintre două variabile, X_i și X_j, și se calculează astfel:

$$a_{ij} = \frac{t_{ij}}{\sqrt{t_{ii}\cdot t_{jj}}},$$ unde t_{ij} este termenul general al matricei $T = R^{-1}$.

Corelația parțială reprezintă legătura liniară dintre două variabile în condițiile în care sunt neutralizate efectele celorlalte variabile din model asupra celor două variabile.

Interpretarea valorilor KMO:

[0.90,1.00] - Foarte bună factorabilitate

[0.80, 0.9) - Bună factorabilitate

[0.70 ,0.8) - Medie

[0.60, 0.7) - Mediocră

[0.50, 0.6) - Slabă

[0.00, 0.5) - Fără factori