mwp-105

factor - 因子分析　【 評価版 】

`factor` は因子分析の機能を提供するコマンドです。`factor` の場合は変数の形で与えられたデータを処理します。これに対し、相関行列や共分散行列の形で与えられたデータを分析するための `factormat` というコマンドも別に用意されています。

1. 因子分析

因子分析 (factor analysis)、より厳密には探索的因子分析 (exploratory factor analysis) というのはデータ削減 (data reduction) のための統計手法です。p 個の変数によって規定されたデータがあった場合、それらの線形結合によって構成される q 個（ただし $q < p$ とする）の新変数で情報の多くが表現でき、しかもそれらの新変数の意味がそれなりに解釈が可能なときに、因子分析という手法は有用性を発揮することになります。

もともとの変数値を $y_{i1}, y_{i2}, \ldots, y_{ip}$ としたとき、因子分析は q 個の共通因子 (common factors) $z_{i1}, z_{i2}, \ldots, z_{iq}$ を算出し、$y_{ij}\,(j = 1, \ldots, p)$ を

$$y_{ij} = \sum_{k=1}^{q} z_{ik} b_{kj} + e_{ij} \tag{1}$$

のような $z_{ik}\,(k = 1, \ldots, q)$ の線形結合として表現します。この式において係数値の集合である $b_{kj}\,(k = 1, \ldots, q;\ j = 1, \ldots, p)$ は因子負荷量 (factor loadings)、残差に相当する $e_{ij}\,(j = 1, \ldots, p)$ は変数 j の独自因子 (unique factor) と呼ばれます。z_{ik}, b_{kj}, e_{ij} の値は一意に決まるわけではありませんが、種々の制約を設けることによって特定の値が算出されることになります。

これらの因子と負荷量が推定できると、次に必要となるのが解釈 (interpretations) のステップです。これは多分に主観的なプロセスであるわけですが、通常は因子負荷量 b_{kj} を評価し、それぞれの共通因子に適切な名前を付けるというステップを踏みます。その際、因子負荷量を回転させるという操作を伴うこともあります。回転には直交回転 (orthogonal rotations) と斜交回転 (oblique rotations) の 2 種類があります。直交回転の場合にはオリジナルの解の性質が維持されるのに対して、斜交回転の場合には一部の性質が失われます。回転のパターンは無限にあるため、同一のデータに対してさまざまな解釈がもたらされることになります。回転については [MV] **factor postestimation** (*mwp-106*) または [MV] **rotate** (*mwp-109*) をご参照ください。

構造方程式モデリング (structural equation modeling) の枠組みを使うとより一般的な形で因子分析が行えます（[SEM] **Intro 5** (*mwp-359*), [SEM] **Example 1** (*mwp-123*) 他参照)。

2. factor コマンドの用例

▷ Example 1: 主因子法

[MV] **factor** おける用例中ではすべて Example データセット bg2.dta が使用されています。

```
. use http://www.stata-press.com/data/r16/bg2.dta [*1]
(Physician-cost data)
```

このデータセットは医師の費用に関する姿勢を記録したものです。具体的には 568 人の医師に対して次のような 6 つの質問を投げかけ、その回答が“賛成”1 から“反対”5 までの 5 段階評価で記録されています。

変数名	意味
bg2cost1	最良の健康管理には高い費用がかかる
bg2cost2	費用は重要な考慮点である
bg2cost3	検査に要する費用を最初に決める
bg2cost4	起り得る合併症のみを監視する
bg2cost5	費用によらずあらゆる手段を講じる
bg2cost6	不要な検査であってもやらないよりましである

このデータセットはオリジナルのものとは異なります。オリジナルと同一の相関行列を持つように corr2data コマンド（[D] **corr2data** 参照）を使って用意されたものです。

[*1] メニュー操作：File ▷ Example Datasets ▷ Stata 16 manual datasets と操作、Multivariate Statistics Reference Manual [MV] の factor の項よりダウンロードする。

ただし評点は平均が 0、標準偏差が 1 となるような形で正規化されている点に注意してください。

```
. summarize bg2cost*, separator(0)
```
[*2]

```
. summarize bg2cost*, separator(0)

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    bg2cost1 |        568    9.45e-09           1  -3.097306   3.057153
    bg2cost2 |        568    6.99e-09           1  -3.651067   3.157189
    bg2cost3 |        568   -6.98e-09           1   -3.20276   3.456272
    bg2cost4 |        568   -1.11e-08           1   -3.07254   2.769688
    bg2cost5 |        568    3.34e-10           1  -3.487679   3.428148
    bg2cost6 |        568    7.86e-09           1  -2.864862   3.011781
```

実際、先頭から 5 人の医師についてのデータをリスト出力してみると次のようになっています。

```
. list bg2cost* in 1/5
```
[*3]

```
     +-------------------------------------------------------------------------------+
     | bg2cost1   bg2cost2   bg2cost3   bg2cost4   bg2cost5   bg2cost6 |
     |-------------------------------------------------------------------------------|
  1. | -1.915584   .9380358  -.2946705   .3302429  -1.427679  -1.012556 |
  2. |  .3637919   -.575019  -1.503126   1.150729  -.0272486  -.9664596 |
  3. |  2.203013  -.3559501  -.5612639  -1.680151  -.2462112  -.4184681 |
  4. |  .4410569  -.3932109   .5441247  -.7309039  -.4729931  -.3337976 |
  5. | -.7046851  -.2237654  -.0054742   .6152834  -1.508267  -1.036314 |
     +-------------------------------------------------------------------------------+
```

このデータセットを対象に `factor` コマンドを実行してみます。最初はデフォルトである主因子法 (principal-factor method) を用いることにします。

- Statistics ▷ Multivariate analysis ▷ Factor and principal component analysis ▷ Factor analysis と操作
- Model タブ: Variables: `bg2cost1-bg2cost6`

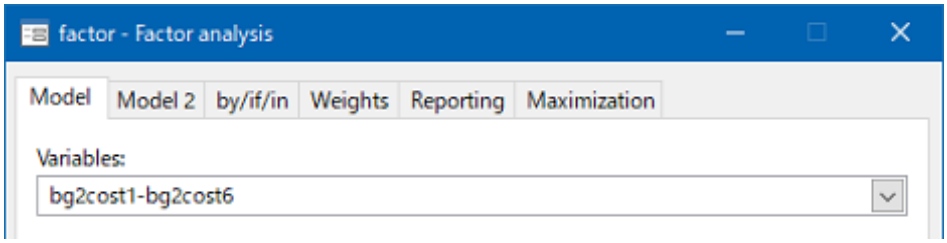

図 1 `factor` ダイアログ- Model タブ

[*2] メニュー操作： Statistics ▷ Summaries, tables and tests ▷ Summary and descriptive statistics ▷ Summary statistics
[*3] メニュー操作： Data ▷ Describe data ▷ List data

- Model 2 タブ: Method: Principal factor （デフォルト）

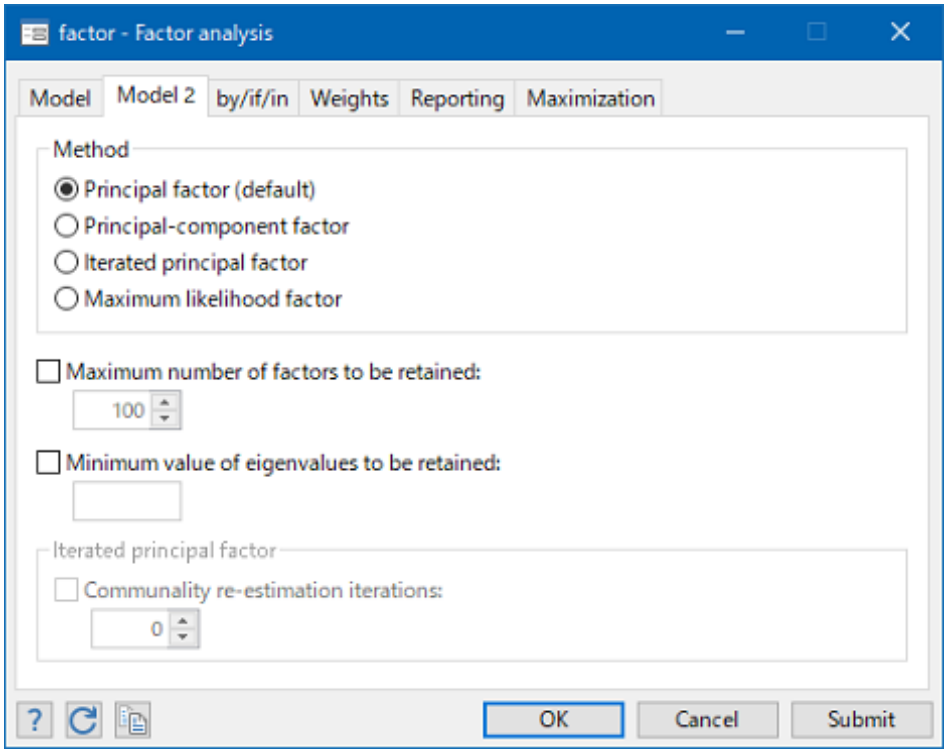

図 2 `factor` ダイアログ- Model 2 タブ

```
. factor bg2cost1-bg2cost6
(obs=568)

Factor analysis/correlation                      Number of obs    =        568
    Method: principal factors                    Retained factors =          3
    Rotation: (unrotated)                        Number of params =         15

    ------------------------------------------------------------------------------
        Factor  |  Eigenvalue   Difference        Proportion   Cumulative
    ------------+-----------------------------------------------------------------
       Factor1  |      0.85389      0.31282            1.0310       1.0310
       Factor2  |      0.54107      0.51786            0.6533       1.6844
       Factor3  |      0.02321      0.17288            0.0280       1.7124
       Factor4  |     -0.14967      0.03951           -0.1807       1.5317
       Factor5  |     -0.18918      0.06197           -0.2284       1.3033
       Factor6  |     -0.25115            .           -0.3033       1.0000
    ------------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(15) =  269.07 Prob>chi2 = 0.0000
```

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Uniqueness
bg2cost1	0.2470	0.3670	-0.0446	0.8023
bg2cost2	-0.3374	0.3321	-0.0772	0.7699
bg2cost3	-0.3764	0.3756	0.0204	0.7169
bg2cost4	-0.3221	0.1942	0.1034	0.8479
bg2cost5	0.4550	0.2479	0.0641	0.7274
bg2cost6	0.4760	0.2364	-0.0068	0.7175

factor からの出力中、第 2 のテーブルには因子負荷量 b_{kj} の値が表示されているわけですが、それからもわかるように共通因子としては Factor1 から Factor3 までの 3 つが選択されています。これは第 1 のテーブルに示されている固有値の中から正のもののみを抽出するというロジックの結果であるわけですが、この閾値については mineigen() オプションによって調整できます。主因子法の場合、5×10^{-6}（実質上 0）がデフォルト値として設定されています。この例の場合、Factor3 に対する固有値は 0 に近い値であるため、意味のある共通因子は Factor1 と Factor2 の 2 個のみと考えられます。

一方、第 2 のテーブル中の Uniqueness というのは独自性を意味し、共通因子では説明できない分散のパーセンテージを表しています。例えば変数 bg2cost1 の場合について言えば

$$1 - 0.2470^2 - 0.3670^2 - 0.0446^2 = 0.8023$$

のようにして算出されています。実際、電卓機能を用いて計算してみると

```
. display 1 - 0.2470^2 - 0.3670^2 - 0.0446^2
.80231284
```

[*4]

となることが確認できます。bg2cost2-bg2cost6 についても同様です。

この例の場合、独自性の値はいずれも高い値を示しているため、共通因子だけでは変数 bg2cost1-bg2cost6 を十分に説明しきれていないと考えなくてはなりません。それでも強いて解釈を試みるなら次のようになるでしょう。

- Factor1 は費用に対する医師の平均的な姿勢を表すものと考えられます。なぜなら因子負荷量テーブルの第 1 列の符合が示すように、すべての質問について“positive”に応答するよう働きかける効果を有するからです。なお、bg2cost2-bg2cost4 についての符合が負であるにもかかわらず“positive”であると主張している点に注意してください。bg2cost2-bg2cost4 については回答の向きが逆になっているからです。費用は治療に対して大きな影響を持つべきではないと医師が考えるなら、その医師は bg2cost2-bg2cost4 に対して反対の立場を取り、残りの 3 つの質問に対しては同意することが予想されます。

[*4] メニュー操作： Data ▷ Other utilities ▷ Hand calculator

- Factor2 はすべての項目について正の符合を有しています。従ってもっともらしいと受け取れるアイデアであればその内容によらず合意してしまう傾向を表すものと解釈することができます。この共通因子は統計学的に見ると残すべきと言えますが、主観的な立場からすると落したくなるかも知れません。

評価版では割愛しています。

▷ Example 2: altdivisor **オプション**

評価版では割愛しています。

▷ Example 3: **主成分因子法**

評価版では割愛しています。

▷ Example 4: **反復主因子法**

評価版では割愛しています。

▷ Example 5: **最尤法**

評価版では割愛しています。

■