社会统计学及SPSS软件应用
STATISTICS WITH SPSS

Instructor:王荣欣

Email: rxwang@qq.com

周一3-4节、单周周四3-4节，3A106-2

2020年11月16日

CONTENTS

- 多因子模型（Multiple-factor model）
- 因子载荷

CONTENTS

很多观测变量之间的高度相关是由某些共同的潜在特性所导致，用因子代表更为本质的潜在特征。

因子分析的应用：

1. 寻求基本结构（summarization）
2. 数据化简（data reduction）

因子分析：起源

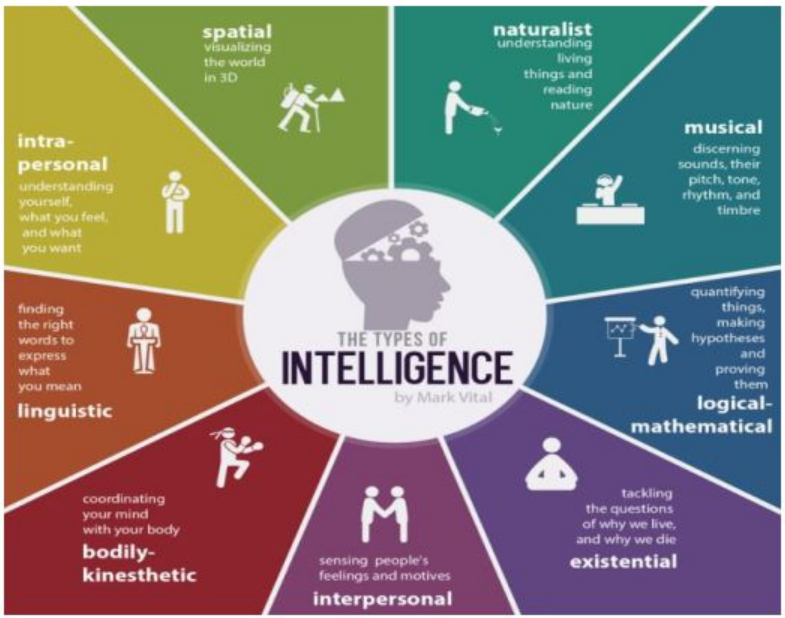
1904年，英国心理学家Charles Spearman研究了33名学生在古典语、法语和英语三门语言课成绩的表现，发现这三门课成绩的相关系数矩阵为：

$$\mathbf{R} = \begin{matrix} \text{Classics} \\ \text{French} \\ \text{English} \end{matrix} \begin{pmatrix} 1.00 & & \\ 0.83 & 1.00 & \\ 0.78 & 0.67 & 1.00 \end{pmatrix}$$

- 为什么这三门课中彼此 **相关系数** 都很高？
- 是否是这三门课成绩的背后是由一个共同的　因素来决定的——**语言能力** ？
- **单因子模型** ——很多相关性很高的变量背后都是由一个公共因子驱动，即每个变量都可以粗略地由这个公共因子表示

因子模型：潜变量

- 共同因子通常是观测不到的 **潜变量**，如智力、社会阶层、满意度、理解力等等，都是研究者很感兴趣，却无法通过测量直接得到的变量。这在心理学、社会学、语言学、经济学等领域非常常见。
- 通常一个公共因子是不够的，错综复杂的变量可能需要多个公共因子才能刻画，这就是更为多见的多因子模型。



1. 数学、物理、化学这三门课成绩之间非常相关，而语文、历史、英语彼此也很相关。而这两组学科跨组的相关性就没有这么高了。
2. 所以，我们直觉上就会觉得，这六门课的成绩会不会是由两个公共因子驱动的，其中一个主要解释前三门，另一个主要解释后三门呢？
3. 有此想法，我们就可以建立如下这种“有两个公共因子存在”的多因子模型：

数学成绩=a_1公共因子1+b_1公共因子2+数学特殊因子

物理成绩=a_2公共因子1+b_2公共因子2+物理特殊因子

化学成绩=a_3公共因子1+b_3公共因子2+化学特殊因子

语文成绩=a_4公共因子1+b_4公共因子2+语文特殊因子

历史成绩=a_5公共因子1+b_5公共因子2+历史特殊因子

英语成绩=a_6公共因子1+b_6公共因子2+英语特殊因子

模型长什么样？

$$X_1 = a_{11}F_1 + \ldots + a_{1d}F_d$$
$$X_2 = a_{21}F_1 + \ldots + a_{2d}F_d$$
$$\ldots\ldots$$
$$X_p = a_{p1}F_1 + \ldots + a_{pd}F_d$$

原始变量：X，
维度是p

因子：F，维度是d

- d远小于p，起到降维目的
- 因子观测不到，是潜在的

系数：a，未知参数需要求解
系数矩阵又称“因子载荷”矩阵

因子分析的数学模型

该模型可用矩阵表示为：

$$\boldsymbol{X} = \boldsymbol{AF} + \boldsymbol{\varepsilon}$$

即

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

其中，所有因子载荷组成的矩阵称作因子载荷矩阵，记为A。

1 系数a_1，a_2，a_3，a_4，a_5，a_6就是公共因子1对六科成绩分别的解释力，叫做因子载荷（factor loading）。

（1） 统计上，其实就是该因子和相应变量之间的相关性。

2 $f_1, f_2, \ldots, f_m$ 叫作公因子，是各个观测变量所共有的因子，解释了变量之间的相关。

3 u_i称为特殊因子，它是每个观测变量所特有的因子，相当于多元回归中的残差项，表示该变量不能被公因子所解释的部分。

因子载荷阵的统计意义

因子载荷

$$\begin{aligned} Cov(X_i, F_j) &= Cov(\sum_{k=1}^{m} a_{ik}F_k + \varepsilon_i, F_j) \\ &= Cov(\sum_{k=1}^{m} a_{ik}F_k, F_j) + Cov(\varepsilon_i, F_j) \\ &= a_{ij} \end{aligned}$$

若X已经过标准化处理，则

$$r_{X_i,F_j} = \frac{Cov(X_i, F_j)}{\sqrt{D(X_i)}\sqrt{D(F_j)}} = Cov(X_i, F_j) = a_{ij}$$

即因子载荷a_{ij}是X_i和F_j的相关系数，它一方面表示了X_i对F_j的依赖程度，另一方面也反映了X_i对F_j的相对重要性。

1 共性方差（communality）：所有公共因子对变量X_i的方差所作的贡献。

(1) 因子载荷矩阵A中第i行元素的平方和，即$h_i^2 = a_{i1}^2 + a_{i2}^2 \cdots + a_{im}^2$。

2 个性方差

3 方差贡献：某个因子所解释的总方差，表示该公共因子对所有变量方差贡献的总和。

(1) 因子载荷矩阵A中第j列元素的平方和。

通过变量之间的相关关系，找到几个基本能刻画这些变量的共同的因素。

下面有一个量表，表中的0代表我国根本不存在这方面的问题，10代表这方面的问题在我国非常严重，请您在量表中选择一个数字表示您的态度

- 环保
- 教育
- 国防
- 稳定
- 民主
- 耕地
- 言论
- 腐败
- 出版
- 犯罪
- 酗酒

因子分析的步骤

1 计算所有变量的相关矩阵。
   （1） 大部分相关系数大于0.3。
   （2） KMO测度。

2 提取因子（求解初始因子）。

3 进行因子旋转（通过坐标变换使因子解的实际意义更容易解释）。

4 因子的解释与命名。

5 计算因子值。

提取因子（因子载荷矩阵的求解）

1. 主成分分析法（Principal Component Factors）
2. 主因子法（Principle Factors）
3. 迭代公因子方差的主因子法（Iterated Principle Factors）
4. 最大似然因子法（Maximum Likelihood Factors）

提取因子：主成分分析法

主成分法是从原始变量的总体方差变异出发，尽可能使其能够被公因子（主成分）所解释，并且使得各公因子对原始变量的方差变异的解释比例依次降低。

- 因子个数的确定
  - 求特征值
  - 做碎石图

求特征值

$$AX = \lambda X$$

- 向量X称为矩阵A的特征向量，λ称为矩阵A的**特征值**。

$$AX = \lambda X = \lambda EX$$

$$(A - \lambda E)X = 0$$

求特征值、因子旋转

1. 对这些指标做因子分析，结果显示从中可提取N个特征值（Eigenvalue）超过1的因子，即所说的成分（component）这N个因子是相互关联的，有交集，有重合。
2. 可是，我们为了便于分析，需要把各个因子关系密切的指标鉴别出来，然后用这些指标建构量表。这时，就需要旋转这N个因子。
3. 因子分析默认各个因子相互关联，旋转因子就是让它们彼此分离，从而更清晰地呈现因子结构。

因子旋转：正交旋转

1. 最常用的旋转方法就是对因子进行直角旋转，就是假定它们零相关。
2. 方差最大化（varimax），即maximize the amount of variance each factor accounts for，让每个因子解释的指标的方差达到最大程度。
3. Stata默认采取最大方差正交旋转。
4. 旋转以后，再看哪几个指标构成一个因子。

正交旋转

- 所有旋转方法的目标都是为了得到尽可能简单的因子结构。
- 正交旋转：假定因子之间不相关。
  1. 四次方最大旋转法：专注于简化因子矩阵的行
  2. 最大方差旋转法：专注于简化因子矩阵的列
  3. 均等变化法

因子得分

因子得分即样本在公共因子上的相应取值。由于公共因子个数往往小于变量个数，所以我们无法通过矩阵变换求得各样本的因子得分。常见的估计因子得分的方法有加权最小二乘法和回归法。

加权最小二乘法是将因子模型看作典型的多元回归模型，此时由于特殊方差σ_i^2是不相等的，所以该回归模型具有异方差性，可以采取加权最小二乘法对参数进行估计，因子得分F即为我们需要估计的参数。

回归法也称汤姆森回归，该方法将公共因子对原始变量作回归，即

$$\hat{F}_j = b_{j1}X_1 + b_{j2}X_2 + \cdots + b_{jp}X_p = \boldsymbol{BX}, \quad j = 1, 2, \cdots, m$$

STATA 命令

```
1  correlate x1 x2 x3 x4 x5
2  factor x1 x2 x3 x4 x5, pcf
3  rotate 或者rotate,varimax
4  loadingplot, factors(2) yline(0) xline(0) 因子载荷图
5  predict f1 f2 因子得分系数矩阵
6  list x1 f1 f2
7  correlate f1 f2
8  scoreplot, mlabel(x1) yline(0) xline(0) 因子得分示意图
9  estat kmo
10 screeplot
11 alpha x1 x2 x3 x4 x5
```

参考文献

1 方匡南，2018，《数据科学》，北京：电子工业出版社。

2 李连江，2017，《戏说统计：文科生的量化方法》，北京：中国政法大学出版社。

3 潘蕊等，2018，《数据思维实践：从零经验到数据英才》，北京：北京大学出版社。