合理进行多元分析——证实性因子分析

胡纯严[1],胡良平[1,2*]

(1. 军事科学院研究生院,北京 100850;
2. 世界中医药学会联合会临床科研统计学专业委员会,北京 100029
*通信作者:胡良平,E-mail:lphu927@163.com)

【摘要】 本文目的是介绍与证实性因子分析有关的基本概念、计算方法、一个实例以及使用SAS实现计算的方法。基本概念包括显变量与潜变量、内生变量与外生变量、探索性与证实性、拟合优度、前提条件;计算方法涉及基本原理和数学模型、证实性因子模型的计算;一个实例的资料是"邓阜仙岩体样品的部分化学成分及其含量";借助SAS软件,对实例中的数据进行了证实性因子分析,并对SAS输出结果做出了解释。

【关键词】 内生变量;外生变量;因子结构;证实性因子分析;样品聚类分析
中图分类号:R195.1 文献标识码:A doi:10.11886/scjsws20230925002

Reasonably carry out multivariate analysis: confirmatory factor analysis

Hu Chunyan[1], *Hu Liangping*[1,2*]

(1. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*
*Corresponding author: *Hu Liangping*, *E-mail*: *lphu*927@163.*com*)

【Abstract】 The purpose of this paper was to introduce the basic concepts, calculation methods, one example related to the confirmatory factor analysis, and how to implement calculations using SAS. Basic concepts included manifest variables and latent variables, endogenous variables and exogenous variables, exploratory and confirmatory, goodness of fit and prerequisites. The calculation methods involved the calculation of basic principles, mathematical models and confirmatory factor models. The data in the example was "partial chemical composition and content of Dengfuxian rock mass samples". With the help of SAS software, the confirmatory factor analysis was conducted on the data in the example and an explanation was given to the SAS output results.

【Keywords】 Endogenous variables; Exogenous variables; Factor structure; Confirmatory factor analysis; Sample cluster analysis

证实性因子分析是一种能够核实公因子的结构以及间接实现样品聚类分析的多元统计分析方法,也被称为确定性因子分析,它是在探索性因子分析的基础上进行的,进一步分析潜在因子和指标之间已经确定的关系,以及潜在因子之间的相关程度。本文将介绍与证实性因子分析有关的基本概念、计算方法、一个实例以及使用SAS实现计算的方法,并对SAS输出结果做出解释。

1 基本概念

1.1 显变量与潜变量

可以直接观测其取值的变量被称为显变量;不可直接观测取值但在背后支配或控制着显变量取值的变量被称为潜变量,也被称为潜在因子。

1.2 内生变量与外生变量

内生变量是指在一个设定的模型中,受其他变量影响或被其他变量解释的变量,可看作模型或方程中的因变量;外生变量是指在一个设定的模型中,只影响或解释其他变量,而不被其他变量影响或解释的变量,可看作模型或方程中的自变量。当内生变量既被其他变量影响、同时又影响其他变量时,该内生变量也可被称为中介变量[1-2]。

1.3 探索性与证实性

在所研究问题中,若对潜变量的数量和重要性一无所知时,只能提出假定的因子模型,通过实际观测到的数据进行试探性建模,并对因子模型中的载荷进行估计和检验,在统计学上,将这样一个实现因子分析的过程称为探索性因子分析。基于探索性因子分析的初步结果,可以知道在所研究问题中,一共存在多少个可能重要的潜变量,也可以知道哪些潜变量对哪些显变量具有不可忽视的控制作用。此时,最好在原有资料的基础上增大样本,针对探索性因子分析得到的"精炼因子模型"重新构建因子模型,并对因子模型进行模型鉴别、参数估计和假设检验、模型的拟合优度检验,以进一步确认所获得的因子模型。这个过程被称为证实性因子分析[3-4]。

1.4 拟合优度

所建立的因子模型对资料的拟合效果，可通过两种方法进行评价。其一，计算拟合优度统计量，例如赤池信息准则、施瓦茨贝叶斯准则等；其二，进行拟合优度检验，例如χ^2检验。

1.5 前提条件

对于证实性因子分析而言，假设的前提条件和探索性因子分析是不完全一样的。首先，探索性因子分析模型要求潜在因子是相互独立的，因为它的主要目的是从众多指标中寻找尽可能少的具有代表性的潜在因子，如果不独立，这些潜在因子就没有代表性；而证实性因子分析模型不要求潜在因子相互独立，因为它的主要目的是估计潜在因子之间的相关性。其次，探索性因子分析模型要求潜在因子ξ_j是方差为1的变量；而证实性因子分析模型不要求潜在因子ξ_j是方差为1的变量。

这两种因子分析模型共同的前提条件：①要求X_i是随机变量；②要求误差项δ_i是互相独立的、均值为0、方差为常数的正态随机变量；③要求误差项δ_i与所有潜在因子ξ_j相互独立。对于证实性因子分析模型的矩阵形式而言，其前提条件可表述为：X是随机向量；δ是均值向量为零向量、方差向量为常向量的多元正态随机向量；ξ与δ相互独立。

2 计算方法

2.1 基本原理和数学模型

从理论上看，证实性因子分析和探索性因子分析的数学模型相同，其表达见式(1)。

$$X_i = a_{i1}\xi_1 + a_{i2}\xi_2 + \cdots + a_{iq}\xi_q + \delta_i,\ i = 1,2,\cdots,k \quad (1)$$

式(1)中，$X_1, X_2, \cdots, X_k$是k个可测变量(或称为显变量)；$\xi_1, \xi_2, \cdots, \xi_q$是$q$个潜在因子(或称为隐变量)，$q \leqslant k$；$a_{ij}$($j$=1,2,...,q)是待估计的因子载荷；$\delta_i$是误差项。

证实性因子分析和探索性因子分析的区别：进行探索性因子分析时，总是假定研究者对指标的内在结构以及隐含的潜在因子一无所知，或知之甚少，因此，估计模型(1)中的未知因子载荷a_{ij}(j=1,2,...,q)时，所有的因子载荷都应估计，也就是说，探索性因子分析是一种非限制性的分析，其结果完全取决于已知数据；而证实性因子分析是在探索性因子分析的基础上进行的，它不需要估计所有的因子载荷，只需估计特定的因子载荷，其余因子载荷均假定为零[5-6]。

2.2 证实性因子分析的计算

2.2.1 由样本导出的与模型隐含的方差协方差矩阵

假定证实性因子分析模型(1)中含有k个观测变量(X变量)。令S是由原始数据计算出来的关于这k个观测变量的方差协方差矩阵，由于它不受任何条件的限制，故称之为非限制性方差协方差矩阵，也称为由样本导出的方差协方差矩阵。又假设证实性因子分析模型(1)的三个基本矩阵a、Φ和θ中共含有p个未知元素，称为待估计的未知参数或自由参数，其余的元素为0，称为固定参数。令ω是由这p个未知参数构成的向量，它是一个p维向量。对于ω的任意一组给定的值，都可以由模型(1)计算出X的一组预测值。将由这组X的预测值计算出来的方差协方差矩阵记为$\Sigma(\omega)$，则有式(2)。

$$\begin{aligned}\Sigma(\omega) &= E[(a\xi + \delta)(a\xi + \delta)^T] \\ &= E(a\xi\xi^T a^T) + E(a\xi\delta^T) + E(\delta\xi^T a^T) + E(\delta\delta^T) \\ &= aE(\xi\xi^T)a^T + aE(\xi\delta^T) + E(\delta\xi^T)a^T + E(\delta\delta^T) \quad (2)\end{aligned}$$

由于有式(3)(假设的前提条件)成立。

$$E(\xi\delta^T)=E(\delta\xi^T)=0,\ E(\xi\xi^T)=\Phi,\ E(\delta\delta^T)=\theta \quad (3)$$

将式(3)带入式(2)，则有式(4)成立。

$$\Sigma(\omega)=a\Phi a^T + \theta \quad (4)$$

这就是说，由一个特定的证实性因子分析模型导出的方差协方差矩阵$\Sigma(\omega)$可以表示成它的三个基本矩阵的函数。一般称$\Sigma(\omega)$为模型(1)隐含的方差协方差矩阵，也称为拟合的方差协方差矩阵[7]。

由以上分析可知，由样本导出的方差协方差矩阵S是由观测数据计算得到的，它是一个与参数ω无关的k阶方阵，表示原始变量之间的相关程度。模型隐含的方差协方差矩阵$\Sigma(\omega)$是由拟合模型的预测值计算出来的，它是一个与参数ω有关的k阶方阵，表示预测变量之间的相关程度。因此，如果一个模型可以很好地描述变量之间的关系，则这两个矩阵的对应元素应当很接近。证实性因子分析的统计检验就是以此为依据的。

2.2.2 未知参数的估计和检验

对于一个特定的证实性因子分析模型(1)，估计其未知参数的方法有很多，例如最大似然估计法、广义最小二乘法、非加权最小二乘法、加权最小二乘法等。这些方法都要求数据符合正态分布。下面简要介绍迭代法估计参数的过程。

令ω_0是模型(1)中参数的一个初始估计，S是由原始观测变量导出的方差协方差矩阵，$\Sigma(\omega_0)$是由这个初始估计确定的模型所隐含的方差协方差矩

阵。令式(5)是一个由模型(1)导出的关于参数ω的迭代公式。

$$\omega_i = g(\omega_{i-1}),\quad i=1,2,\cdots \tag{5}$$

如果$\Sigma(\omega_0)$与S无限接近,即达到预先给定的精确度要求,则ω_0可以被认为是模型中未知参数ω的一个“准确”的估计。如果$\Sigma(\omega_0)$与S的接近程度未达到预先给定的精确度要求,则将初始估计ω_0代入迭代式(5),计算出新的估计ω_1,然后检查ω_1是否满足要求。如果仍不满足要求,则将ω_1代入迭代式(5),计算出又一个新的估计ω_2,如此迭代下去,直到求出一个满足要求的ω的估计,并称最后得到的这个估计为模型(1)的一个非标准参数估计。求出非标准参数估计值后,分别对每一个估计值计算其标准误和t值,并进行统计检验。此处,标准误的估计是以渐近分布理论为依据的。

2.2.3　计算标准因子载荷

与多重线性回归分析中标准回归系数的估计类似,在证实性因子分析中,也可以计算因子载荷的标准估计值。有两种方法:一是将观测变量标准化,即用观测变量的相关系数矩阵进行分析;二是利用公式计算,例如,指标X_i在潜在因子ξ_j上的因子载荷a_{ij}的标准估计值可用式(6)计算出来。

$$a_{ij}^{(s)}=a_{ij}S_{\xi_j}/S_{X_i} \tag{6}$$

式(6)中,$S_{\xi j}$和S_{Xi}分别是潜在因子ξ_j和指标X_i的标准差。标准因子载荷消除了量纲的影响,可用来比较各指标对潜在因子的相对重要性。绝对值越大,指标对潜在因子的贡献越大。

3　实例与SAS实现

3.1　问题与数据结构

3.1.1　实际问题及数据

【例1】邓阜仙岩体样品的部分化学成分及其含量见表1[8]。试对此资料进行证实性因子分析。对表1资料的探索性因子分析结果表明,有一个因子支配着X_1,记为ξ_1,称为非金属氧化物因子;另一个因子支配着X_2、X_3、X_4和X_5,记为ξ_2,称为金属氧化物因子。证实性因子分析路径图见图1。

【例2】沿用例1资料,如果原始数据丢失,但知道对原始数据加工后得到的相关系数矩阵,见表2,其他条件不变,请在此基础上对资料进行证实性因子分析。

表1　邓阜仙岩体样品的部分化学成分及其含量

Table 1　Partial chemical composition and content of Dengfuxian rock mass samples

样　品	$SiO_2(X_1)$	$TiO_2(X_2)$	$FeO(X_3)$	$CaO(X_4)$	$K_2O(X_5)$
1	75.20%	0.14%	1.86%	0.91%	5.21%
2	75.15%	0.16%	2.11%	0.74%	4.93%
3	72.19%	0.13%	1.52%	0.69%	4.65%
4	72.35%	0.13%	1.37%	0.83%	4.87%
5	72.74%	0.10%	1.41%	0.72%	4.99%
6	73.29%	0.03%	1.07%	0.17%	3.15%
7	73.72%	0.03%	0.77%	0.28%	2.78%

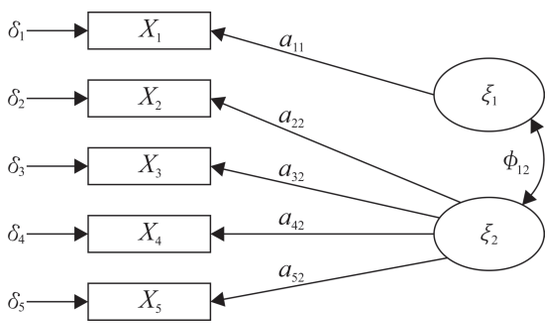

图1　例1资料对应的证实性因子分析路径图

Figure 1　Confirmatory factor analysis path map for example 1 data

表2　相关系数矩阵

Table 2　Correlation coefficient matrix

变　量	X_1	X_2	X_3	X_4	X_5
X_1	1.000				
X_2	0.239	1.000			
X_3	0.518	0.910	1.000		
X_4	0.143	0.922	0.766	1.000	
X_5	0.116	0.923	0.835	0.960	1.000

3.1.2　对数据结构的分析

例1资料是一个单组设计5元定量资料。例2资料是例1资料的另一种表达形式,本质上仍是单组设计5元定量资料。

3.2　用SAS实现统计分析

3.2.1　分析例1的资料

设所需要的SAS程序如下[9]:

```
data a1;
input id$ X1-X5;
cards;
1 75.20 0.140 1.86 0.91 5.21
2 75.15 0.160 2.11 0.74 4.93
3 72.19 0.130 1.52 0.69 4.65
4 72.35 0.130 1.37 0.83 4.87
5 72.74 0.100 1.41 0.72 4.99
6 73.29 0.033 1.07 0.17 3.15
```

```
7 73.72 0.033 0.77 0.28 2.78
;
run;
proc calis data=a1 method=lsml;
lineqs X1= f1+e1,
X2= f2+e2,
X3=a32 f2+e3,
X4=a42 f2+e4,
X5=a52 f2+e5;
std e1-e5=var_e1-var_e5, f1 f2=2*var_f;
cov f1 f2=cov;
run;
quit;
```

【SAS主要输出结果及解释】线性方程中，标准回归系数的估计结果见表3。因子模型中5个参数与0之间差异均有统计学意义，说明通过探索性因子分析得到的因子模型基本成立。

表3 线性方程中的标准化效应

Table 3 Standardization effects in linear equations

变 量	预测变量	参 数	估 计	标准误	t	Pr>|t|
X_1	f_1	-	0.039	0.017	2.362	0.018
X_2	f_2	-	0.952	0.044	21.871	<0.001
X_3	f_2	a_{32}	0.853	0.115	7.402	<0.001
X_4	f_2	a_{42}	0.970	0.032	30.679	<0.001
X_5	f_2	a_{52}	0.981	0.025	38.784	<0.001

外生变量方差的标准化结果的估计值见表4。5个外生变量中只有第一个e_1具有统计学意义，而其他4个均无统计学意义。严格地说，可将其他4个外生变量设置为0，重新运行SAS程序，以获得因子模型中其他参数更稳定的估计结果。

表4 外生变量方差的标准化结果

Table 4 Normalized results of the variance of exogenous variables

变量类型	变 量	参 数	估 计	标准误	t	Pr>|t|
误差	e_1	var_e_1	0.998	0.001	763.200	<0.001
	e_2	var_e_2	0.094	0.083	1.128	0.260
	e_3	var_e_3	0.272	0.197	1.385	0.166
	e_4	var_e_4	0.059	0.061	0.965	0.335
	e_5	var_e_5	0.037	0.050	0.747	0.455
隐藏	f_1	var_f_1	1.000	-	-	-
	f_2	var_f_2	1.000	-	-	-

由表3和表4的计算结果，可写出线性方程组中各方程的具体表达式，见式(7)。

$$\begin{cases} X_1 = 0.039f_1 + 0.998e_1 \\ X_2 = 0.952f_2 + 0.094e_2 \\ X_3 = 0.853f_2 + 0.272e_3 \\ X_4 = 0.970f_2 + 0.059e_4 \\ X_5 = 0.981f_2 + 0.037e_5 \end{cases} \quad (7)$$

外生变量中协方差的标准化结果的估计值见表5。两个隐变量之间的协方差的估计值为4.439，此值与0之间差异无统计学意义，即可以认为两个隐变量互相独立。

表5 外生变量中协方差的标准化结果

Table 5 Normalized results of covariance in exogenous variables

Var1	Var2	参 数	估 计	标准误	t	Pr>|t|
f_1	f_2	cov	4.439	10.338	0.429	0.668

基于上述表3、表4、表5的结果，绘制出例1资料对应的证实性因子分析路径图，见图2。

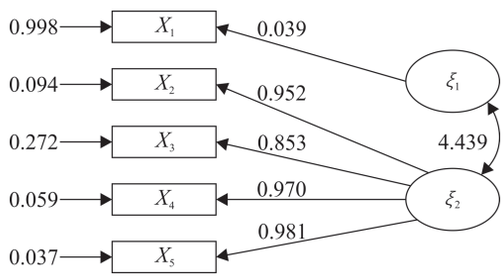

图2 带有参数估计值的例1资料对应的证实性因子分析路径图

Figure 2 Confirmatory factor analysis path map for example 1 data with parameter estimation values

3.2.2 分析例2的资料

设所需要的SAS程序如下[9]：

```
data a2(type=corr);
infile cards missover;
_type_='corr';
if _n_=1 then _type_='n';
input _name_ $ X1-X5;
CARDS;
n 7 7 7 7 7
X1 1.00000 0.23918 0.51767 0.14322 0.11554
X2 0.23918 1.00000 0.91026 0.92200 0.92308
X3 0.51767 0.91026 1.00000 0.76621 0.83477
X4 0.14322 0.92200 0.76621 1.00000 0.95980
X5 0.11554 0.92308 0.83477 0.95980 1.00000
;
RUN;
proc calis data=a2 method=lsml;
lineqs X1= f1+e1,
X2= f2+e2,
X3=a32 f2+e3,
```

```
X4=a42 f2+e4,
X5=a52 f2+e5;
std e1-e5=var_e1-var_e5,f1 f2=2*var_f;
cov f1 f2=cov;
run;
quit;
```

【SAS主要输出结果及解释】与例1的输出结果完全相同(因为是同一个资料,只是输入的数据格式不同而已),此处从略。

4 讨论与小结

4.1 讨论

4.1.1 潜在因子的尺度问题

任何一个观测变量都是有尺度的,即有原点和单位。何为潜在因子的尺度?例如,在例1中,有一个因子支配着X_1,记为ξ_1,称为非金属氧化物因子;另一个因子支配着X_2、X_3、X_4和X_5,记为ξ_2,称为金属氧化物因子。ξ_1和ξ_2都是不可直接观测的理论变量或潜在因子,它们显然是没有尺度的,即没有原点和单位。为了使得潜在因子之间具有可比性,必须对每一个潜在因子定义它的原点和单位。假设潜在因子的均值为0,也就是说,只要将原始数据标准化,潜在因子的原点问题就得以解决。在此基础上,解决潜在因子的单位问题有两个常用方法:一是假定所有潜在因子的方差为1,即假定潜在因子的单位等于样本所代表的总体的标准差;另一个最常用的、最方便的方法是在每一个潜在因子所支配的几个观测变量中,选择一个作为参照变量,并假定该潜在因子对这个参照变量的影响为1,即参照变量在这个因子上的因子载荷为1,这就意味着该潜在因子的尺度被假定为与参照变量的尺度一样。

4.1.2 模型的可鉴别性

设模型中观测变量总个数为k,并令$c=k(k+1)/2$,称为模型的信息量;又设模型中待估计的未知参数的个数为p,则有下列命题成立:对于可鉴别的模型(包括正好可鉴别和过分可鉴别两种情形),一定有$c\geqslant p$。也就是说,如果$c<p$,模型一定是不可鉴别的。如果$c<p$,解决这个问题的一种办法是增加限制条件:如限定某些变量的因子载荷为1,则未知参数的个数就减少2;还可限定某个指标的度量误差为0或近似为0,则未知参数的个数将进一步减少。于是,原来不可鉴别的模型就可能转化为可鉴别的模型。此外,还可以通过增加观测变量解决不可鉴别的问题。

4.2 小结

本文介绍了与证实性因子分析有关的基本概念、计算方法、一个实例以及使用SAS实现计算的方法。基本概念包括显变量与潜变量、内生变量与外生变量、探索性与证实性、拟合优度、前提条件;计算方法涉及基本原理和数学模型、证实性因子模型的计算;一个实例的资料是“邓阜仙岩体样品的部分化学成分及其含量”;借助SAS软件,对实例中的数据分别进行了证实性因子分析,并对SAS输出结果做出了解释。

参考文献

[1] 张岩波.潜变量分析[M].北京:高等教育出版社,2009:35-59.
Zhang YB. Latent variables analysis[M]. Beijing: Higher Education Press, 2009: 35-59.

[2] Johnson RA, Wichern DW. 实用多元统计分析[M].6版.北京:清华大学出版社,2008:481-538.
Johnson RA, Wichern DW. Applied multivariate statistical analysis[M]. 6th edition. Beijing: Tsinghua University Press, 2008: 481-538.

[3] Armitage P, Colton T. Encyclopedia of biostatistics[M]. 2nd edition. New York: John Wiley & Sons, Inc, 2005: 2053-2072.

[4] Lattin JM, Carroll JD, Green PE. 多元统计分析[M].北京:机械工业出版社,2003:171-205.
Lattin JM, Carroll JD, Green PE. Analyzing multivariate data[M]. Beijing: China Machine Press, 2003: 171-205.

[5] 李卫东.应用多元统计分析[M].北京:北京大学出版社,2008:207-238.
Li WD. Applied multivariate statistical analysis[M]. Beijing: Peking University Press, 2008: 207-238.

[6] 余锦华,杨维权.多元统计分析与应用[M].广州:中山大学出版社,2005:210-231.
Yu JH, Yang WQ. Multivariate statistical analysis and application[M]. Guangzhou: Sun Yat-sen University Press, 2005: 210-231.

[7] 张润楚.多元统计分析[M].北京:科学出版社,2006:190-217.
Zhang RC. Multivariate statistical analysis[M]. Beijing: Science Press, 2006: 190-217.

[8] 胡良平.面向问题的统计学:(3)试验设计与多元统计分析[M].北京:人民卫生出版社,2012:139-164.
Hu LP. Problem-oriented statistics: (3) experimental design and multivariate statistical analysis[M]. Beijing: People's Medical Publishing House, 2012: 139-164.

[9] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 2715-2824.

(收稿日期:2023-09-25)
(本文编辑:陈 霞)