**34.因子分析**

**（一）基本原理**

**一、概述**

因子分析，是用少数起根本作用、相互独立、易于解释通常又是不可观察的因子来概括和描述数据，表达一组相互关联的变量。通常情况下，这些相关因素并不能直观观测。

因子分析是从研究相关系数矩阵内部的依赖关系出发，把一些具有错综复杂关系的变量归结为少数几个综合因子的一种多变量统计分析方法。简言之，即用少数不可观测的隐变量来解释原始变量之间的相关性或协方差关系。

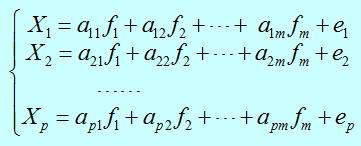
因子分析的作用是减少变量个数，根据原始变量的信息进行重组，能反映原有变量大部分的信息；原始部分变量之间多存在较显著的相关关系，重组变量（因子变量）之间相互独立；因子变量具有命名解释性，即该变量是对某些原始变量信息的综合和反映。

主成分分析是因子分析的特例。主成份分析的目标是降维，而因子分析的目标是找出公共因素及特有因素，即公共因子与特殊因子。

因子分析模型在形式上与线性回归模型相似，但两者有着本质的区别：回归模型中的自变量是可观测到的，而因子模型中的各公因子是不可观测的隐变量，而且两个模型的参数意义也不相同。

**二、原理**

假设样品检测*p*个指标（变量）*X1*, …, *Xp*，得到观测矩阵*X*，这*p*个指标变量可能受*m*(*m<p*)个共同因素*f1,…fm*的影响，再加上其它影响因素。表示为：



用矩阵表示为



其中，共同影响因素*f1,…fm*是均值为0方差为1的随机变量，称为公共因子；*Ap×m*称为因子载荷矩阵，*aji*是第*j*变量在第*i*公共因子上的负荷，即*Xj*在坐标轴*fi*上的投影；*ei*是变量*Xi*所特有的因子，均值为0方差为*σi2*，称为特殊因子。各特殊因子之间及特殊因子与公共因子之间都是相互独立的，即COV(*ei, ej*)=0, COV(*e, f*)=0.

因子分析就是用*f1,…fm*代替*X1, …, Xp*, 达到降维的目的。

主成份分析中，残差通常是彼此相关的。因子分析中，特殊因子起到残差的作用，但被定义为彼此不相关且和公因子也不相关。而且每个公因子假定至少对两个变量有贡献，否则它将是一个特殊因子。

在开始提取公共因子时，已假定它们彼此不相关且具有单位方差。故向量*X*的协方差矩阵Σ可以表为

*Σ=D(X)=D(Af+e)=AAT+D*

其中，D=diag(*σ12*, …, *σp2*). 若假定*X*已标准化，即每个*Xi*都均值为0方差为1. 则



记称为变量共同度，则有.

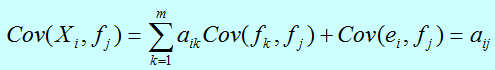
*hi2*反映了公共因子*f*对*Xi*的影响（贡献），即*Xi*对共同因素*f*的依赖程度；*σi2*为剩余方差，若*hi2*接近1，*σi2*很小，则表明因子分析的效果好。

公共因子*fj*对*Xi*的影响，可由*A*中第j列元素来描述，



称为公共因子*fj*对*X*的贡献，是衡量公共因子重要性的尺度。

*aij*为第*i*个变量与第*j*个公共因子的相关系数，反映了它们的相关程度：



**三、求因子载荷矩阵**

若*X*的协方差矩阵∑和*D*已知，则根据∑-*D*=*AAT*求出*A*



其中，λ1≥λ2≥…λm>λm+1=…=λp=0为∑\*=∑*-D*的特征值，*ξi*为相应的特征向量。

但在实际问题中，并不知道∑和*D*，就需要从*n*个样品，*p*个指标的*np*个样本数据，估计因子载荷*aij*和特殊因子方差*σi2*. 常用的参数估计法有：主成分法、主因子解法、极大似然法。

（1）主成分法

记样本数据的协方差矩阵为，其特征值λ1≥λ2≥…≥λp≥0, 相应的特征向量为*ξi*. 当最后*p-m*个特征值较小时，对做谱分解：

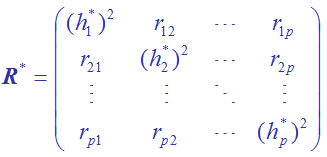


先取，看是否接近对角矩阵，若是，说明只取一个公共因子就行了，所有指标主要受到这一个公共因子的影响；若否，再取，看是否接近对角矩阵……象主成分分析一样，直接取前*q*个特征值和特征向量，使得它们的特征值之和占全部特征值之和的85％以上即可。此时，特殊因子方差为



（2）主因子解法

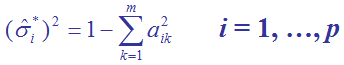
主因子解法是对主成分法的修正，记样品数据的相关矩阵为R，设*R= AAT + D*，则*R\* = R–D = AAT*称为约相关矩阵。若已知特殊因子方差的初始估计，则，故



计算*R\**的特征值λ1\*≥λ2\*≥…≥λp\*>0，相应的特征向量为u*i*\*, 则有近似分解式：*R\** = AAT, 其中



令



则A和D为因子模型的一个解，称为主因子解。

在实际中特殊因子方差是未知的，以上得到的解是近似解。为了得到近似程度更好的解，常常采用迭代主因子法。即利用上面得到的



作为特殊因子方差的初始估计，重复上述步骤，直到解稳定为止。

变量共同度*hi2*常用的初始估计有以下三种方法：

① 取第*i*个变量与其他所有变量的多重相关系数的平方；

② 取第*i*个变量与其他变量相关系数绝对值的最大值；

③ 取1（等价于主成分解）。

（3）极大似然法

假定公共因子*f*和特殊因子*e*服从正态分布，则可得到因子载荷阵和特殊因子方差的极大似然估计。设*p*维观测向量X(1)，...，X(n)为来自正态总体*Np*(*μ,*∑)的随机样品，则样品似然函数为*μ*，∑的函数L(*μ*,∑).

设∑= AAT +D，取，则似然函数为A，D的函数：φ(A,D)，求A，D使φ(A,D)达到最大。为保证得到唯一解，可附加计算上方便的唯一性条件：ATD-1A = 对角阵，用迭代方法可求得极大似然估计A和D。

**四、公共因子的解释**

得到估计的因子模型后，还必须对得到的公因子进行解释。即对每个公共因子给出一种意义明确的名称，用来反映在预测每个可观察变量中这个公因子的重要性。该公因子的重要程度就是在因子模型矩阵中相应于这个因子的系数，显然这个因子的系数绝对值越大越重要，而接近0则表示对可观察变量没有什么影响。

因子解释是一种主观的方法，有时侯通过旋转公因子可以减少这种主观性，也就是要使用非奇异的线性变换。

1. 因子旋转

将因子载荷矩阵*A*右乘一个正交矩阵*T*后得到一个新的矩阵*A\**。它并不影响变量*Xi*的共同度*hi*2，却会改变因子的方差贡献*gj2*。因子旋转通过改变坐标轴，能够重新分配各个因子解释原始变量方差的比例，使因子更易于理解。

设*p*维可观测向量*X*满足因子模型：*X = AF +e*。*T*为正交矩阵，则因子模型可写为

*X = ATTTF + e = A\*F\* + e*

其中*A\* = AT，F\* = TTF*. 易知，

∑= AAT + D = A\*A\*T + D

其中A\* = AT. 这说明，若A，D是一个因子解，任给正交阵T，A\* = AT，D也是因子解。在这个意义下，因子解是不惟一的。

由于因子载荷阵不惟一，故可对因子载荷阵进行旋转。目的是使因子载荷阵的结构简化，使载荷矩阵每列或行的元素平方值向0和1两极分化，这样的因子便于解释和命名。

有三种主要的正交旋转法：四次方最大法、方差最大法和等量最大法。这些旋转方法的目标是一致的，只是策略不同。

如果两种旋转模型导出不同的解释，这两种解释不能认为是矛盾的。倒不如说是看待相同事物的两种不同方法，是在公因子空间中的两个不同点。在统计意义上所有旋转都是一样的，即不能说一些旋转比另一些旋转好。因此，在不同的旋转方法之间进行的选择必须根据非统计观点，通常选择最容易解释的旋转模型。

2. 因子得分

计算因子得分，是用原有变量来描述因子，第*j*个因子在第*i*个样品上的值可表示为：

*Fji* = *βj*1*xi*1 + *βj*2*xi*2 +…+ *βjpxip* *j* = 1, …, *k*

其中，*xi*1, …, *xip*分别是第1, …, *p*个原有变量在第*i*个样品上的取值，*βj*1, …, *βjp*分别是第*j*个因子和第1, …, *k*个原有变量间的因子值系数。可见，它是原有变量线性组合的结果(与因子分析的数学模型正好相反)，因子得分可看作各变量值的加权(*βj*1, …, *βjp*)总和，权数的大小表示了变量对因子的重要程度。

于是有

*Fj* = *βj*1*X*1+*βj*2*X*2+…+*βjpXp* *j* = 1, …, *k*

称为因子得分函数。由于因子个数*k*小于原有变量个数*p*，故式中方程的个数少于变量的个数。因此，对因子值系数通常采用最小二乘意义下的回归法进行估计。可将上式看作是因子变量*Fj*对*p*个原有变量的线性回归方程(其中常数项为0)。可以证明，式中回归系数的最小二乘估计满足：*Bj* = *Aj*T*R*-1，其中

*Bj* = (*βj*1,*βj*2,…,*βjp*)， *Aj*T = (*a*1*j*,*a*2*j*,…,*apj*)

为第1, …, *p*个变量在第*j*个因子上的因子载荷，*R*-1为原有变量的相关系数矩阵的逆矩阵。

由上式计算出因子变量*Fj*的因子值系数，再利用因子得分函数可算出第*j*个因子在各个样品上的因子得分。

**（二）SAS实现**

**一、PROC FACTOR过程步**

**基本语法：**

PROC FACTOR DATA=数据集 <可选项> ;

<PRIORS 计算公因子方差的方法;>

VAR变量列表;

<PARTIAL 变量列表;>

<FREQ 变量;>

<WEIGHT变量;>

<BY变量;>

**说明：**1. 可选项：

out＝数据集——存储原始数据和因子得分估计；

outstat＝数据集——用于存储因子分析的结果，可以用来做进一步因子分析；

method=因子选择方法

常用的包括：principal（主成分法），ml（极大似然法），prinit（迭代主因子法），usl（未加权最小二乘因子法），alpha（ 因子法或称harris法）

Heywood

公因子方差大于1时令其为1，并允许迭代继续执行下去。因为公因子方差是相关系数的平方（介于0和1之间），若最终的公因子方差=1，称为Heywood状况；若大于1，称为超-Heywood状况，因子解无效。

priors=计算公因子方差的方法

即给各变量的公因子方差赋初值，包括one（=1.0），max (最大绝对相关系数 )，smc（多元相关系数的平方），asmc (与多元相关系数的平方成比例，但要适当调整使它们的和等于最大绝对相关)。

nfactors=n——指定保留的因子个数；

mineigen=p——指定保留因子的最小特征值；

 proportion=p

使用先验公因子方差估计，对被保留的因子规定所占公共方差比例为这个p值。

converge=p

当公因子方差的最大改变小于p时停止迭代，默认为0.001；

 maxiter=n——规定迭代的最大步数，默认为30；

 rotate＝因子旋转方法

常用的有varimax（正交的最大方差转轴法）、orthomax（由gamma=指定权数的正交方差最大转轴法）和promax（在正交最大方差转轴的基础上进行斜交旋转）。

norm＝kaiser | raw | weight | cov | none

为了对因子模型进行旋转，规定模型矩阵中行的正规化方法。

2. 输出选项

*simple*——输出简单统计数。

*corr*——输出相关阵和偏相关阵。

*score*——输出因子得分模型中的系数。

*scree*——输出特征值的屏幕图。

*ev*——输出特征向量。

*residuals*——输出残差相关阵和有关的偏相关阵。

*msa*——输出被所有其余变量控制的每对变量间的偏相关，并抽样适当的Kaiser度量。

*reorder*——在输出时让各种因子矩阵的这些行重新排序。按因子顺序，绝对载荷大的变量先被输出。

3. PLOTS=绘图选项

ALL——绘制所有图形；

INITLOADINGS/ LOADINGS/PRELOADINGS——未旋转的/旋转的/预先旋转的因子载荷，子选项有CIRCLE=, FLIP, NPLOTS=, PLOTREF, VECTOR；

SCREE——陡坡图和变量解释，子选项有UNPACK；

4. priors语句

为var变量设定公因子方差，值在0.0和1.0之间。其值的设定应与var语句的变量相对应。

**二、PROC SCORE过程步**

无论是初始因子模型还是旋转后的因子模型，都是将指标表示为公因子的线性组合。在因子分析中，还可以将公因子表示为指标的线性组合，这样就可以从指标的观测值估计各个公因子的值（因子得分）。它对样品的分类有实际意义。因子得分可由proc score过程完成，可以输出特征值情况、因子载荷、公因子解释比例等。

**基本语法：**

PROC SCORE DATA = <原始数据集>;

VAR <用来计算得分的原始变量列表>;

或者在FACTOR过程步中加上可选项SCORE。

注：factor和score两个过程写同一程序可以提高效率。

**例1**对全国30个省市自治区经济发展基本情况的八项指标作因子分析。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 省份 | GDP  x1 | 居民消费水平x2 | 固定资产投资x3 | 职工平均工资x4 | 货物周转量  x5 | 居民消费价格指数  x6 | 商品零售价格指数x7 | 工业总产值  x8 |
| 北京 | 1394.89 | 2505 | 519.01 | 8144 | 373.9 | 117.3 | 112.6 | 843.43 |
| 天津 | 920.11 | 2720 | 345.46 | 6501 | 342.8 | 115.2 | 110.6 | 582.51 |
| 河北 | 2849.52 | 1258 | 704.87 | 4839 | 2033.3 | 115.2 | 115.8 | 1234.85 |
| 山西 | 1092.48 | 1250 | 290.9 | 4721 | 717.3 | 116.9 | 115.6 | 697.25 |
| 内蒙古 | 832.88 | 1387 | 250.23 | 4134 | 781.7 | 117.5 | 116.8 | 419.39 |
| 辽宁 | 2793.37 | 2397 | 387.99 | 4911 | 1371.1 | 116.1 | 114 | 1840.55 |
| 吉林 | 1129.2 | 1872 | 320.45 | 4430 | 497.4 | 115.2 | 114.2 | 762.47 |
| 黑龙江 | 2014.53 | 2334 | 435.73 | 4145 | 824.8 | 116.1 | 114.3 | 1240.37 |
| 上海 | 2462.57 | 5343 | 996.48 | 9279 | 207.4 | 118.7 | 113 | 1642.95 |
| 江苏 | 5155.25 | 1926 | 1434.95 | 5943 | 1025.5 | 115.8 | 114.3 | 2026.64 |
| 浙江 | 3524.79 | 2249 | 1006.39 | 6619 | 754.4 | 116.6 | 113.5 | 916.59 |
| 安徽 | 2003.58 | 1254 | 474 | 4609 | 908.3 | 114.8 | 112.7 | 824.14 |
| 福建 | 2160.52 | 2320 | 553.97 | 5857 | 609.3 | 115.2 | 114.4 | 433.67 |
| 江西 | 1205.11 | 1182 | 282.84 | 4211 | 411.7 | 116.9 | 115.9 | 571.84 |
| 山东 | 5002.34 | 1527 | 1229.55 | 5145 | 1196.6 | 117.6 | 114.2 | 2207.69 |
| 河南 | 3002.74 | 1034 | 670.35 | 4344 | 1574.4 | 116.5 | 114.9 | 1367.92 |
| 湖北 | 2391.42 | 1527 | 571.68 | 4685 | 849 | 120 | 116.6 | 1220.72 |
| 湖南 | 2195.7 | 1408 | 422.61 | 4797 | 1011.8 | 119 | 115.5 | 843.83 |
| 广东 | 5381.72 | 2699 | 1639.83 | 8250 | 656.5 | 114 | 111.6 | 1396.35 |
| 广西 | 1606.15 | 1314 | 382.59 | 5105 | 556 | 118.4 | 116.4 | 554.97 |
| 海南 | 364.17 | 1814 | 198.35 | 5340 | 232.1 | 113.5 | 111.3 | 64.33 |
| 四川 | 3534 | 1261 | 822.54 | 4645 | 902.3 | 118.5 | 117 | 1431.81 |
| 贵州 | 630.07 | 942 | 150.84 | 4475 | 301.1 | 121.4 | 117.2 | 324.72 |
| 云南 | 1206.68 | 1261 | 334 | 5149 | 310.4 | 121.3 | 118.1 | 716.65 |
| 西藏 | 55.98 | 1110 | 17.87 | 7382 | 4.2 | 117.3 | 114.9 | 5.57 |
| 陕西 | 1000.03 | 1208 | 300.27 | 4396 | 500.9 | 119 | 117 | 600.98 |
| 甘肃 | 553.35 | 1007 | 114.81 | 5493 | 507 | 119.8 | 116.5 | 468.79 |
| 青海 | 165.31 | 1445 | 47.76 | 5753 | 61.6 | 118 | 116.3 | 105.8 |
| 宁夏 | 169.75 | 1355 | 61.98 | 5079 | 121.8 | 117.1 | 115.3 | 114.4 |
| 新疆 | 834.57 | 1469 | 376.95 | 5348 | 339 | 119.7 | 116.7 | 428.76 |

**代码：**

**data** Economics;

infile 'C:\MyRawData\Economics.txt';

input Province $ x1 x2 x3 x4 x5 x6 x7 x8;

**run**;

**proc** **factor** data = Economics n=**3** SIMPLE CORR plots=(scree);

var x1-x8;

**run**;

**运行结果及说明：**

| **30 个观测的均值和标准差** | | |
| --- | --- | --- |
| **变量** | **均值** | **标准差** |
| **x1** | 1921.0927 | 1474.8060 |
| **x2** | 1745.9333 | 861.6419 |
| **x3** | 511.5083 | 402.8855 |
| **x4** | 5457.6333 | 1310.2181 |
| **x5** | 666.1200 | 459.9353 |
| **x6** | 117.2867 | 2.0253 |
| **x7** | 114.9067 | 1.8981 |
| **x8** | 862.9980 | 584.5873 |

| **相关性** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **x1** | **x2** | **x3** | **x4** | **x5** | **x6** | **x7** | **x8** |
| **x1** | 1.00000 | 0.26676 | 0.95058 | 0.19058 | 0.61724 | -0.27256 | -0.26363 | 0.87374 |
| **x2** | 0.26676 | 1.00000 | 0.42614 | 0.71813 | -0.15101 | -0.23514 | -0.59273 | 0.36310 |
| **x3** | 0.95058 | 0.42614 | 1.00000 | 0.39963 | 0.43062 | -0.28049 | -0.35905 | 0.79186 |
| **x4** | 0.19058 | 0.71813 | 0.39963 | 1.00000 | -0.35559 | -0.13496 | -0.53916 | 0.10438 |
| **x5** | 0.61724 | -0.15101 | 0.43062 | -0.35559 | 1.00000 | -0.25318 | 0.02172 | 0.65858 |
| **x6** | -0.27256 | -0.23514 | -0.28049 | -0.13496 | -0.25318 | 1.00000 | 0.76284 | -0.12522 |
| **x7** | -0.26363 | -0.59273 | -0.35905 | -0.53916 | 0.02172 | 0.76284 | 1.00000 | -0.19207 |
| **x8** | 0.87374 | 0.36310 | 0.79186 | 0.10438 | 0.65858 | -0.12522 | -0.19207 | 1.00000 |

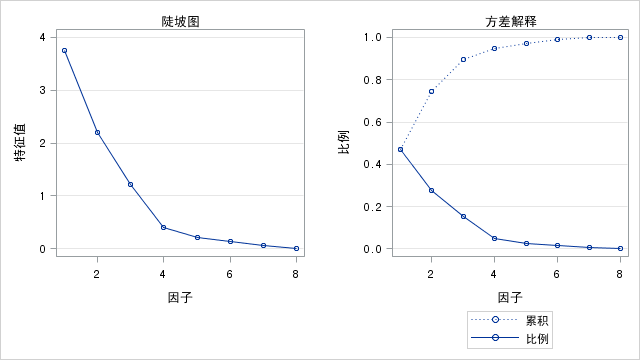
相关系数矩阵

先验公因子方差估计: ONE

| **相关矩阵的特征值: 总计 = 8 平均值 = 1** | | | | |
| --- | --- | --- | --- | --- |
|  | **特征值** | **差值** | **比例** | **累积** |
| **1** | 3.75512808 | 1.55839208 | 0.4694 | 0.4694 |
| **2** | 2.19673600 | 0.98184422 | 0.2746 | 0.7440 |
| **3** | 1.21489177 | 0.81249198 | 0.1519 | 0.8958 |
| **4** | 0.40239979 | 0.18959738 | 0.0503 | 0.9461 |
| **5** | 0.21280241 | 0.07484352 | 0.0266 | 0.9727 |
| **6** | 0.13795890 | 0.07250109 | 0.0172 | 0.9900 |
| **7** | 0.06545781 | 0.05083257 | 0.0082 | 0.9982 |
| **8** | 0.01462524 |  | 0.0018 | 1.0000 |

相关矩阵的特征值和累积贡献，前三个主成分的累积贡献为89.58%，故考虑选取三个公共因子。

3 个因子将被 NFACTOR 准则保留。



绘图选项scree绘制陡坡图

| **因子模式** | | | |
| --- | --- | --- | --- |
|  | **Factor1** | **Factor2** | **Factor3** |
| **x1** | 0.88490 | 0.38362 | 0.12088 |
| **x2** | 0.60671 | -0.59819 | 0.27132 |
| **x3** | 0.91169 | 0.16112 | 0.21198 |
| **x4** | 0.46622 | -0.72241 | 0.36793 |
| **x5** | 0.48583 | 0.73829 | -0.27524 |
| **x6** | -0.50856 | 0.25191 | 0.79664 |
| **x7** | -0.61959 | 0.59438 | 0.43756 |
| **x8** | 0.82273 | 0.42672 | 0.21098 |

因子模型（因子载荷矩阵），即公因子用原始变量表示的回归系数。由因子载荷系数看出，三个因子的含义不易解释，故需要考虑作因子旋转。

| **每个因子说明的方差** | | |
| --- | --- | --- |
| **Factor1** | **Factor2** | **Factor3** |
| 3.7551281 | 2.1967360 | 1.2148918 |

| **最终的公因子方差估计: 总计 = 7.166756** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **x1** | **x2** | **x3** | **x4** | **x5** | **x6** | **x7** | **x8** |
| 0.94482954 | 0.79954948 | 0.90208112 | 0.87460860 | 0.85685809 | 0.95672136 | 0.92862683 | 0.90348083 |

公因子对原始变量的解释能力（方差贡献）的量度分别为3.7551281 2.1967360 1.2148918，三个公因子对变量的解释能力的总和为7.166756. 各个变量的共同度*hi2*都较大（接近1），可以认为三个公因子可以很好地解释原始变量中的信息。

**例2** 为了更好地解释公共因子**，**继续做**因子旋转——**方差最大正交旋转。

在PROC FACTOR语句中再加上一个ROTATE = PROMAX旋转选项，这样将在得到主成分解后再进行方差最大正交旋转(VARIMAX)，并加了REORDER选项使输出时把原始变量受相同因子影响的放一起。

**代码：**

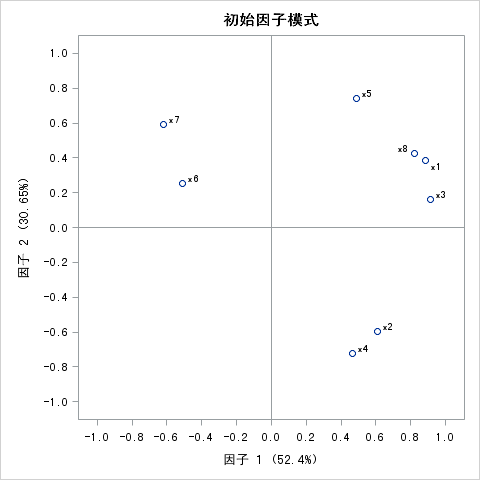
**proc** **factor** data = Economics n=**3** ROTATE = VARIMAX REORDER plots=(initloadings preloadings loadings);

var x1-x8;

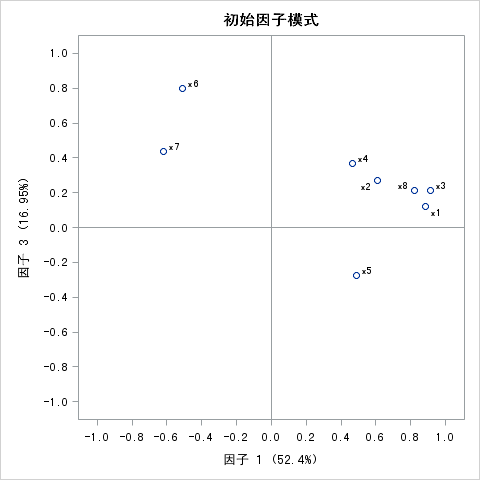
**run**;

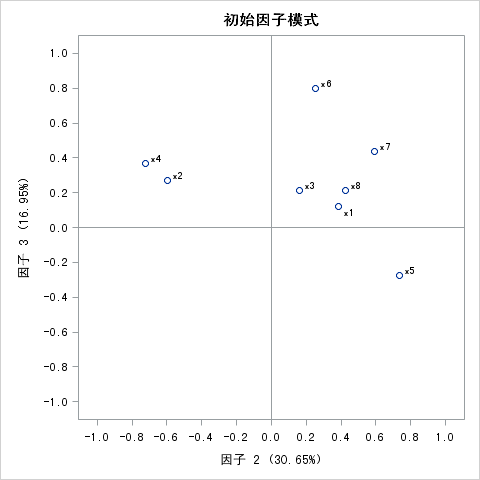
**运行结果及说明：**

初始因子方法: 主成分



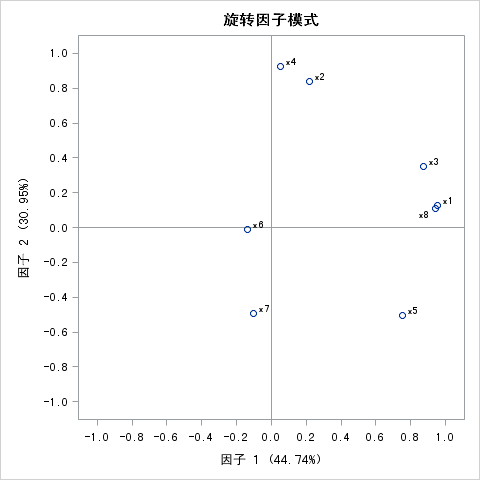
未旋转因子，变量有3个聚点。旋转因子目的是让一些变量取值是0.



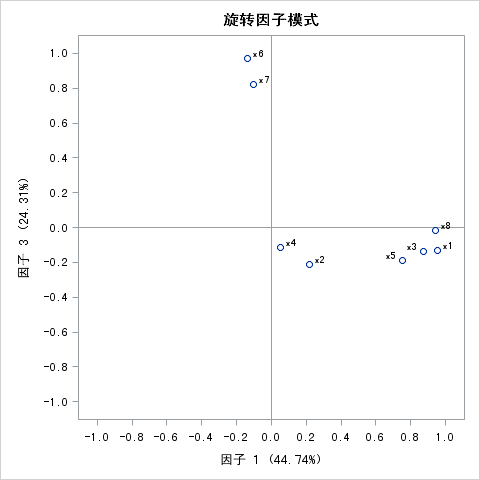


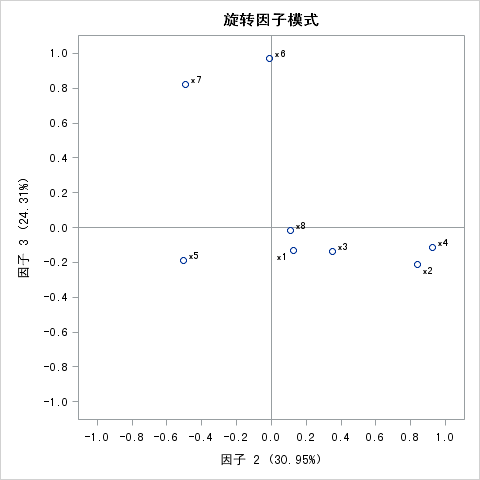
FACTOR 过程

旋转方法: Varimax



旋转后的因子，x2, x4, x6, x7取值接近于0，只剩x1, x3, x5, x8.





旋转方法: Varimax

| **正交变换矩阵** | | | |
| --- | --- | --- | --- |
|  | **1** | **2** | **3** |
| **1** | 0.81731 | 0.40776 | -0.40710 |
| **2** | 0.54776 | -0.76907 | 0.32938 |
| **3** | 0.17878 | 0.49220 | 0.85193 |

| **旋转因子模式** | | | |
| --- | --- | --- | --- |
|  | **Factor1** | **Factor2** | **Factor3** |
| **x1** | 0.95498 | 0.12530 | -0.13091 |
| **x8** | 0.94388 | 0.11114 | -0.01464 |
| **x3** | 0.87129 | 0.35218 | -0.13749 |
| **x5** | 0.75227 | -0.50517 | -0.18909 |
| **x4** | 0.05112 | 0.92679 | -0.11429 |
| **x2** | 0.21672 | 0.84099 | -0.21288 |
| **x6** | -0.13525 | -0.00900 | 0.96868 |
| **x7** | -0.10259 | -0.49440 | 0.82078 |

旋转后的载荷矩阵，表明第一公共因子在GDP （x1）、工业总产值（x8）、固定资产投资（x3）、货物周转量（x5）这些指标上的载荷更加明显，可见该因子反映了经济发展规模；第二个因子在职工平均工资（x4）、居民消费水平（x2）指标上的载荷更加明显，可见该因子反映了消费水平；第三个因子在居民消费价格指数（x6）、商品零售价格指数（x7）指标上载荷更加明显，可见该因子反映了消费价格。

**例3** 进一步可以计算**因子得分**。

在FACTOR过程中使用SCORE选项和OUTSTAT=选项输出得分系数数据集。比如，为了计算方差最大正交旋转的因子得分，并做因子得分散点图。

代码：

**proc** **factor** data = Economics n=**3** ROTATE=VARIMAX REORDER SCORE OUT=Out plots=(scree initloadings preloadings loadings);

**run**;

**proc** **print** data = Out;

var Province factor1 factor2 factor3;

**run**;

**proc** **plot** data = Out;

plot factor2\*factor1 $Province = '\*'/href=**0** vref=**0**;

**run**;

运行结果及说明：

回归估计的评分系数

| **包含每个因子的变量的多重相关系数平方** | | |
| --- | --- | --- |
| **Factor1** | **Factor2** | **Factor3** |
| 1.0000000 | 1.0000000 | 1.0000000 |

| **标准化评分系数** | | | |
| --- | --- | --- | --- |
|  | **Factor1** | **Factor2** | **Factor3** |
| **x1** | 0.30605 | 0.01076 | 0.04635 |
| **x2** | 0.02282 | 0.38523 | 0.03479 |
| **x3** | 0.26980 | 0.12847 | 0.07397 |
| **x4** | -0.02451 | 0.45260 | 0.09915 |
| **x5** | 0.24933 | -0.31723 | -0.13498 |
| **x6** | 0.06936 | 0.17933 | 0.65154 |
| **x7** | 0.07774 | -0.09810 | 0.46312 |
| **x8** | 0.31652 | 0.02542 | 0.12274 |

用回归法得到的因子得分系数，由此可以写出三个因子得分函数：

Factor1 = 0.30605\*x1+0.02282\*x2+0.26980\*x3-0.02451\*x4

+0.24933\*x5+0.06936\*x6+0.07774\*x7+0.31652\*x8

Factor2 = 0.01076\*x1+0.38523\*x2+0.12847\*x3+0.45260\*x4

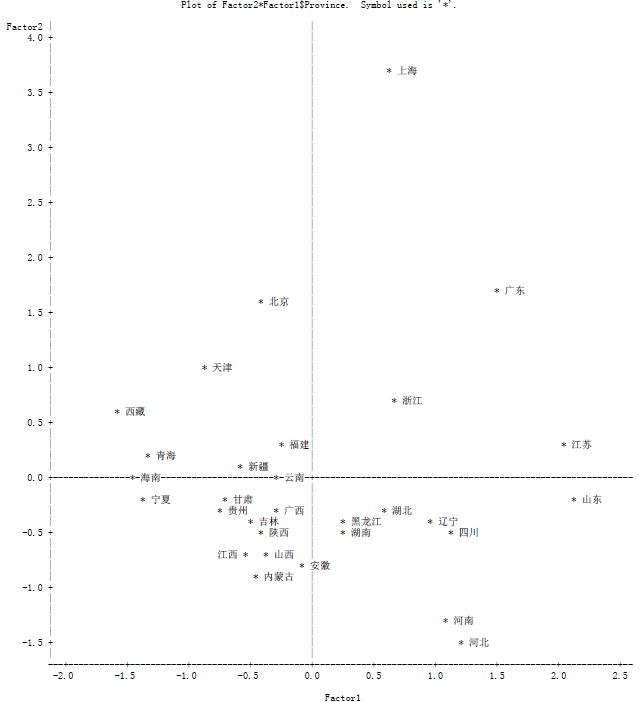
-0.31723\*x5+0.17933\*x6-0.09810\*x7+0.02542\*x8

Factor3 = 0.04635\*x1+0.03479\*x2+0.07397\*x3+0.09915\*x4

-0.13498\*x5+0.65154\*x6+0.46312\*x7+0.12274\*x8

将30个省份的观测值代入因子得分函数，即得各地区的因子得分

| **Obs** | **Province** | **Factor1** | **Factor2** | **Factor3** |
| --- | --- | --- | --- | --- |
| **1** | 北京 | -0.39736 | 1.58700 | -0.25811 |
| **2** | 天津 | -0.88764 | 0.98429 | -1.62974 |
| **3** | 河北 | 1.22842 | -1.52116 | -0.77830 |
| **4** | 山西 | -0.36588 | -0.66516 | -0.14735 |
| **5** | 内蒙古 | -0.47821 | -0.88694 | 0.20671 |
| **6** | 辽宁 | 0.95945 | -0.43272 | -0.61496 |
| **7** | 吉林 | -0.51600 | -0.40157 | -0.94792 |
| **8** | 黑龙江 | 0.23364 | -0.38076 | -0.58367 |
| **9** | 上海 | 0.60483 | 3.66086 | 0.82829 |
| **10** | 江苏 | 2.03433 | 0.26868 | -0.17225 |
| **11** | 浙江 | 0.65154 | 0.74895 | -0.32932 |
| **12** | 安徽 | -0.07044 | -0.79931 | -1.50597 |
| **13** | 福建 | -0.26961 | 0.27185 | -0.79964 |
| **14** | 江西 | -0.56144 | -0.68367 | -0.05008 |
| **15** | 山东 | 2.11774 | -0.19757 | 0.25119 |
| **16** | 河南 | 1.07136 | -1.31827 | -0.46510 |
| **17** | 湖北 | 0.60165 | -0.30001 | 1.26599 |
| **18** | 湖南 | 0.26085 | -0.52386 | 0.51914 |
| **19** | 广东 | 1.48231 | 1.68546 | -1.18361 |
| **20** | 广西 | -0.28371 | -0.27438 | 0.61247 |
| **21** | 海南 | -1.47391 | -0.00569 | -2.25105 |
| **22** | 四川 | 1.10866 | -0.52550 | 0.97792 |
| **23** | 贵州 | -0.76687 | -0.34925 | 1.66330 |
| **24** | 云南 | -0.27804 | 0.04407 | 2.04592 |
| **25** | 西藏 | -1.59335 | 0.63020 | -0.01241 |
| **26** | 陕西 | -0.41400 | -0.53524 | 0.88562 |
| **27** | 甘肃 | -0.71809 | -0.22183 | 1.01825 |
| **28** | 青海 | -1.34460 | 0.18197 | 0.45775 |
| **29** | 宁夏 | -1.35842 | -0.15568 | -0.14352 |
| **30** | 新疆 | -0.57721 | 0.11522 | 1.14042 |



根据前两个因子得分绘制的因子得分散点图。右下部分的山东、辽宁、四川、河南、河北等省份经济发展总量高，消费价格低；右上部分的广东、浙江、江苏等省份的经济发展总量高，消费价格高；左上部分的西藏、清海等地区的经济发展总量低，消费价格高；左下部分的宁夏、甘肃、贵州、江西、内蒙古等省份的经济发展总量低，消费价格低。