变量的相关分析

汪四水

2016.7.16

# 1变量的相关

相关分析是通过计算变量间的相关系数来考察变量间是否具有线性关系，该方法与回归分析的差别就是相关分析时变量间的关系地位是对等的。

由于数据的复杂性，不同类型的数据计算相关程度的量也有所区别，常用的有pearson相关系数，也就是我们常用的相关系数，计算此种相关系数要求变量服从正态分布；一种为Spearman相关系数，该相关系数与变量的分布无关；对于分类数据间的相关关系更多的是使用**Kendall的相关系数**，以及由此引申的出来的**Gamma系数**，**Somers d**系数等，下面将进行说明。

## 1.1两变量的相关

**1.1.1 Pearson 相关系数**

**相关系数的计算** 实践中通过多次观测，便可得到一系列预报量（）与预报因子（）的数据，记为，则

**Pearson矩相关系数：**

**相关系数的显著性检验** 对于一对变量，相关系数的值多大才能称之为这两个变量间有线性相关关系？判断这个问题是否成立，在统计中称之为假设检验，统计假设的提法为一对假设，称为原假设和备择假设，在此问题中其记法为原假设：这两个变量间没有相关关系，即，其备择假设：这两个变量间有相关关系，即。因而这种要检验的假设可简记为。要判断原假设是否成立，寻找到的检验统计量为：，通过查表或者利用SAS软件计算的结果即可判断。

例2.1，数据见例1.1

SAS编程如下：

**data** hb\_hanyang;

input year x y;

label x='3月上旬平均温度（℃）'

y='越冬代发蛾始盛期（6月30日为0）';

datalines;

1961 8.6 3

1962 8.3 5

1963 9.7 3

1964 8.5 1

1965 7.5 4

1966 8.4 4

1967 7.3 5

1968 9.7 2

1969 5.4 7

1970 5.5 5

run;

**proc** **corr** data=hb\_hanyang;

var x y;

**run**;

输出结果如下：

**Pearson** 相关系数, N = 10

当 H0: Rho=0 时，Prob > |r|

x y

x 1.00000 -0.77130

3月上旬平均温度（℃） 0.0090

y -0.77130 1.00000

越冬代发蛾始盛期（6月30日为0） 0.0090

此结果表明，间的相关系数，超过此值的概率值为，此概率值比显著性水平以及都要小，表明间的这种线性相关关系达到了极显著性水平，从而可认为间的确存在线性相关关系。

**1.1.2 Spearman相关系数**

若不服从正态分布，甚至不考虑总体的分布形式，则间的相关关系可采用以下的计算公式：

Spearman的秩相关系数：，其中是在中的秩，是在中的秩，，。

公式中涉及到秩的概念，下面举例说明秩的含义。设给定一组数据，将其从小到大进行排序，则最小的秩为1，第二小的秩为2，依次类推，最大的一个数的秩为。例如给定的一组数为2，-2，5，10，7，3。则按从小到大排序为：-2，2，3，5，7，10，则-2的秩为1，2的秩为2，3的秩为3，5的秩为4，7的秩为5，10的秩为6.。

其检验可通过两种方式达到，一是在小样本的条件下，计算精确的拒绝阈或值，二是在大样本下的近似计算。SAS软件中则可直接计算得到相应的统计量值及值，从而可作出来判断。如：

**proc** **corr** data=hb\_hanyang spearman;

var x y;

**run**;

**Spearman** 相关系数, N = 10

当 H0: Rho=0 时，Prob > |r|

x y

x 1.00000 -0.87007

3月上旬平均温度（℃） 0.0011

y -0.87007 1.00000

越冬代发蛾始盛期（6月30日为0） 0.0011

从输出的结果表明，间的Spearman相关系数，超过此值的概率值为，此概率值比显著性水平以及都要小，表明间的这种线性相关关系达到了极显著性水平，从而可认为间的确存在线性相关关系。

**1.1.2相合系数**

对于分类数据间的相关关系可用**Kendall的**相关系数来度量，在此基础上，还可以定义**Gamma系数**，**Somers d**系数等

设分类变量取有序值分别为：，分类变量取有序值分别为，则其对应的总体确定了一个双向有序列联表，

表2.3.1 表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  | | | | 合计 |
|  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| 合计 | |  |  |  |  |  |

为计算相关系数，令

，

则在值中，值有对，依此类推，值有对，令：

 G称为观测中的和谐对(concordant parirs)数目

 H称为非和谐对（discordant pairs）数目。

则有：

对表而言：



例如（G，H的计算）：在284个陆续住入精神病院的患者中按社会等级和诊断分类得如下频数表，下面来计算G,H这两个量

表2.3.2 社会等级与精神病调查数据

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | 诊断（变量） | | | | 合计 |
|  |  | 神经病 | 抑郁症 | 人格异常 | 精神分裂症 |
| 社会等级  （变量） | 1 | 45 | 25 | 21 | 18 | 109 |
| 2 | 10 | 45 | 24 | 22 | 101 |
| 3 | 17 | 21 | 18 | 18 | 74 |
| 合计 | | 72 | 91 | 63 | 58 | 284 |

G=45(45+24+22+21+18+18)+25(24+22+18+18)+21(22+18) 第一行

+10(21+18+18)+45(18+18)+24（18） 第二行

=12172

从G的计算中可以看出来，它是从表格的左上出发往右，一行一行进行；括号里的数字为划去括号外的元素所在的行列后剩余的表格右下的元素。

H=18（10+45+24+17+21+18）+21（10+45+17+21）+25（10+17） 第一行

+22（17+21+18）+24（17+21）+45（17） 第二行

=7967

从H的计算过程可以看出来，它是从表格的右上出发往左下，一行一行进行；括号里的数字为划去括号外的元素所在的行列后剩下表格中左下的元素。

以下几个系数则是衡量相合性的常用指标：

#### 1.1.2.1 Kendall系数

令,其中，



其值处于-1和1之间，越近于1，越倾向于正相合，反之为负相合。

在SAS系统中输出的是下面的tau-b,tau-c系数。

**** 和分别为两个变量的约束的观测数

由上例可算得此两值如下：

=45（25+21+18）+25（21+18）+21（18） 第一行

+10（45+24+22）+45（24+22）+24（22） 第二行

+17（21+18+18）+21（18+18）+18（18） 第三行

|  |
| --- |
| =14571 |

=45（10+17）+10（17） 第一列

+25（45+21）+45（21） 第二列

+21（24+18）+24（18） 第三列

+18（22+18）+22（18） 第四列

|  |
| --- |
| **=**6410 |

**，其值可达，**

#### 1.1.2.2 Gamma系数

令，此量具有如下的概率解释：随机选择的两个体按两个变量分类时，顺序相同与不同的概率之差。仅当时，；仅当，。

#### 1.1.2.3 Somers d系数

令，此系数常用于或列联表，

注：在度量或列联表的相合性时，**Somers d系数比其它系数更好。**

#### 1.1.2.4相合性的检验

相合性的假设检验问题可以形成如下检验问题：原假设：和相互独立：和相合。

检验统计量为：， 其中为的标准误，其计算公式为：



其近似式为：

在原假设为真时，，则有，利用它可进行假设检验。

**data** cc;

input a$ b$ count@@;

datalines;

1 神经病 45 1 抑郁症 25 1 人格异常 21 1 精神分裂症 18

2 神经病 10 2 抑郁症 45 2 人格异常 24 2 精神分裂症 22

3 神经病 17 3 抑郁症 21 3 人格异常 18 3 精神分裂症 18

**run**;

**proc** **freq**;

tables a\*b/nocum nopercent norow nocol all;

weight count;

**run**;

其输出为：

a b

频数 |1神经病 |2抑郁症 |3人格异硘4精神分羭 合计

--------+--------+--------+--------+--------+

1 | 45 | 25 | 21 | 18 | 109

--------+--------+--------+--------+--------+

2 | 10 | 45 | 24 | 22 | 101

--------+--------+--------+--------+--------+

3 | 17 | 21 | 18 | 18 | 74

--------+--------+--------+--------+--------+

合计 72 91 63 58 284

统计量 值 渐近标准误差

----------------------------------------------------------

Gamma 0.2088 0.0755

Kendall Tau-b 0.1492 0.0545

Stuart Tau-c 0.1564 0.0571

Somers D C|R 0.1584 0.0580

Somers D R|C 0.1405 0.0513

“a \* b”的汇总统计量

Cochran-Mantel-Haenszel 统计量（基于表得分）

统计量 对立假设 自由度 值 概率

----------------------------------------------------------------

1 非零相关 1 **7.0694** 0.0078

2 行均值得分差值 2 11.1010 0.0039

3 一般关联 6 30.6903 <.0001

总样本大小 = 284

**1.1.3 Kappa系数**

#### 1.1.3.1.1Kappa系数

对于行数与列数一样多方格表，其行和列的倾向是否一致？此即所谓的一致性检验。

主对角线元素之和占总数的比例作为一致性的度量，有缺陷，因为在随机时也可能使。因此在方格表的单侧都固定的条件下,我们有:

，

则有，从而得到方格表的一致性系数，Kappa系数如下：



注: 分子为的修正,分母是使得整个值不超过1.

#### 1.1.3.2Kappa系数的检验

在方表双侧都固定的条件下,有如下结论:

，

从而

=

将,,,代入,得:



由即可进行检验

## 1.2一个变量与一组变量的相关（复相关）

变量y与一组变量x1,x2,…,xp间的相关，称为复相关，其定义为，有

若对于变量x1,x2,…,xp，同样考虑xj与其余变量的复相关系数，则可用如下方式计算：

 ，则 ，

检验： 。检验统计量：

## 1.3偏相关

在多变量问题中，任意两个变量的相关，均受到其他变量变化取值的影响，因此称将其他变量固定时的相关为偏相关。

考虑多变量x1,x2,…,xp中，为叙述的方便，考虑x1,x2的偏相关系数：

1. 分别建立x1,x2与x3-xp的线性回归模型，得到残差
2. 计算残差相关系数即偏相关系数

一般地有如下公式：注意有一个负号。



偏相关系数的检验，与回归方程中变量xi前的回归系数的显著性检验一样，即，检验统计量

例三种相关系数的计算：

**data** corr;

input x1-x4 y@@;

datalines;

10 23 3.6 113 15.7 9 20 3.6 106 14.5

10 22 3.7 111 17.5 13 21 3.7 109 22.5

10 22 3.6 110 15.5 10 23 3.5 103 16.9

8 23 3.3 100 8.6 10 24 3.4 114 17

10 20 3.4 104 13.7 10 21 3.4 110 13.4

10 23 3.9 104 20.3 8 21 3.5 109 10.2

6 23 3.2 114 7.4 8 21 3.7 113 11.5

9 22 3.6 105 12.3

**run**;

**proc** **iml**;

use corr;

read all into data;

x=data[,];

nr=nrow(x);

nc=ncol(x);

i=i(nr);

v=repeat(**1**,nr,**1**);

h=i-v\*v`/nr;

sscp=x`\*h\*x;

d=inv(sqrt(diag(sscp)));

r=d\*sscp\*d;

print r;

ir=inv(r);

a=i(nc);

do i=**1** to nc;

do j=**1** to nc;

a[i,j]=-ir[i,j]/sqrt(ir[i,i]\*ir[j,j]);

if i=j then a[i,j]=sqrt(**1**-**1**/ir[i,j]);

if i>j then a[i,j]=r[i,j];

end;

end;

print a;

quit;

用SAS过程直接计算三种系数（需要事先计算哪些变量间的相关，尤其是偏相关方面）

**data** corr;

input x1-x4 y@@;

datalines;

10 23 3.6 113 15.7 9 20 3.6 106 14.5

10 22 3.7 111 17.5 13 21 3.7 109 22.5

10 22 3.6 110 15.5 10 23 3.5 103 16.9

8 23 3.3 100 8.6 10 24 3.4 114 17

10 20 3.4 104 13.7 10 21 3.4 110 13.4

10 23 3.9 104 20.3 8 21 3.5 109 10.2

6 23 3.2 114 7.4 8 21 3.7 113 11.5

9 22 3.6 105 12.3

run;

**proc** **corr** data=corr;

var x1-x4 y;

**run**;

/\*find correlation y and x1,x2,x3,x4,respectively\*/

**proc** **corr** data=corr;

var x1-x4;

with y;

**run**;

/\*fixed variable x4,find correlation between x1,x2,x3\*/

**proc** **corr** data=corr;

var x1-x3;

partial x4;

**run**;

/\*R2 between y and x1-x4\*/

**proc** **reg** data=corr;

model y=x1-x4;

**run**;

## 1.4两组变量的相关（典型相关）

**问题**：寻找两组变量间的相关关系。设

设变量，，且有，，，

注意协方差阵中包含了变量间的关系。

**做法**：找到变量的线性函数（可想象为主成分）---称为**典型变量**，找到变量的线性函数（典型变量），然后再求这两个线性函数的相关系数（此时为简单的pearson相关，即**典型相关系数**），使得相关系数达到最大，然后再找第2个最大的相关系数，如此下去。由于典型变量间互不相关的特点，导致有很好性质。

公式表示：

寻找典型变量：，使得典型变量间的相关系数达到最大，，又因为有 所以限制典型变量的方差为1，故问题表述为：



可用拉格朗日乘子法进行求解（过程略）。

奇异值分解法求典型相关系数和典型变量（寻找不变量）

设变量，（）且有，，，即：

，

由于，因而寻找，的相关问题转变为：协方差阵在非退化变换（即非退化）下的不变量是什么 ：

=，典型相关系数及典型变量就含于矩阵中了，对其进行奇异值分解，即有：

，从变换的过程来看，知，所以有：

典型变量分别为：

典型相关系数为

**结果：**

计算步骤如下：

1. 从式中，求出特征根（此即典型相关系数的平方），并按从大到小排序，以及对应的特征向量（此为典型变量前的系数）
2. 利用已求得的特征根，再从求出特征向量（由此可得典型变量）

注：特征向量间的关系如下：

，两边乘以有：即



即：

典型相关系数的**检验**：。拒绝原假设，则需要进行检验：，一直进行下去。设所需要的检验为：，统计量为：，其中，

性质：

1. 同组的典型变量间不相关（相当于主成分的性质）
2. 不同组相同位次的典型变量间的相关系数为特征根的平方根
3. 原始变量与典型变量间的相关系数

将原始变量，，变为典型变量：，（其变换为， ）

则有，，

所以有：







相关系数：







**data** coresp;

input y$ x1-x3@@;

datalines;

a 12 224 76 b 60 548 39

c 246 659 28

d 416 765 47

**run**;

**proc** **corresp** data=coresp all out=results;

var x1-x3;

id y;

**run**;

**proc** **plot** data=results;

plot dim1\*dim2='\*'$y;

**run**;

**proc** **gplot** data=results;

plot dim1\*dim2;

**run**;

# 2主成分分析

设变量,有，通过坐标变换使得新的分量间不相关（即主成分）

结果：记其中，且，是特征根对应的单位特征向量，因此有

则称为主成分，其中（）称为第主成分。

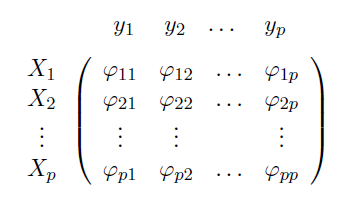
性质： ，因此有：

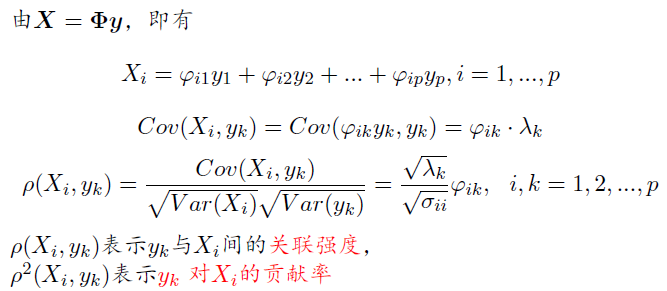
1.（方差没有改变，但主成分已按方差大小进行了排序）

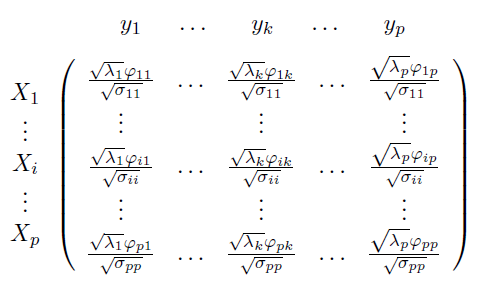
2.主成分间互不相关（从主成分的协方差阵即知）

3.，所以相关系数矩阵为：







 ，所以当变量为标准化变量时，有，注意相关系数的分解。

4提取主成分的个数（达到降维的目的） （可取75%，80%。。。。）

典型相关系数为均为1：



# 3因子分析







则有方差分解式，,由此式可求得因子载荷矩阵A.

**（1）主成分法**



令：

所以有：

从而求得因子得分为：

性质：

（1） 相关系数矩阵为：



即有，

（2）方差分解：



行和：为公共因子方差或共性方差，

列和：为公因子对变量的方差。

因子旋转：对任意正交阵有

，将看成新的因子，仍满足因子分析模型。

方差极大准则：

因子模型：，公因子对变量的方差为：，其中，总平方和为

，将此式中的用替代，有，使之达到极大的方法称为方差极大准则。



正交旋转变换：每次只涉及两个因子，设经旋转后变为因子 则载荷间有关系式：

 ，令，有 ，其中，，，，