（2006年，艾滋病疗法的评价及预测问题）

5对应分析

对应分析（ correspondence analysis ）是在R型和Q型因子分析基础上发展起来的多元统计分析方法，又称为R-Q型因子分析。

例 试用对应分析研究我国部分省份的农村居民家庭人均消费支出结构。选取7个变量：A为食品支出比重，B为衣着支出比重，C为居住支出比重，D为家庭设备及服务支出比重，E为医疗保健支出比重，F为交通和通讯支出比重，G为文教娱乐、日用品及服务支出比重。考察的地区（即样品）有10个：山西、内蒙古、吉林、辽宁、黑龙江、海南、四川、贵州、甘肃、青海（原始数据见表1）。

表1 中国10个省份农村居民家庭人均消费支出数据

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 地区 | A | B | C | D | E | F | G |
| 1山西 | 0.583910 | 0.111480 | 0.092473 | 0.050073 | 0.038193 | 0.018803 | 0.079946 |
| 2内蒙古 | 0.581218 | 0.081315 | 0.112380 | 0.042396 | 0.043280 | 0.040004 | 0.083339 |
| 3辽宁 | 0.565036 | 0.100121 | 0.123970 | 0.041121 | 0.043429 | 0.031328 | 0.078919 |
| 4吉林 | 0.530918 | 0.105360 | 0.116952 | 0.045064 | 0.043735 | 0.038508 | 0.095256 |
| 5黑龙江 | 0.555201 | 0.096500 | 0.143498 | 0.037566 | 0.052111 | 0.026267 | 0.072829 |
| 6海南 | 0.654952 | 0.047852 | 0.095238 | 0.047945 | 0.022134 | 0.018519 | 0.096844 |
| 7四川 | 0.640012 | 0.061680 | 0.116677 | 0.048471 | 0.033529 | 0.017439 | 0.072043 |
| 8贵州 | 0.725239 | 0.056362 | 0.073262 | 0.044388 | 0.016366 | 0.015720 | 0.057261 |
| 9甘肃 | 0.678630 | 0.058043 | 0.088316 | 0.038100 | 0.039794 | 0.015167 | 0.067999 |
| 1青海 | 0.665913 | 0.088508 | 0.096899 | 0.038191 | 0.039275 | 0.019243 | 0.033801 |

数据表41中列变量（A,B,C,D,E,F,G）是消费支出的几个指标，可以理解为属性变量“消费支出”的几个水平（或类目）。表41中的样品（行变量）是几个不同的地区，可理解为属性变量“地区”的几个不同水平（或类目）。

表2和图1给出了计算的主要结果。

表2 惯量和（卡方）分解

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 奇异值 | 主惯量 | 卡方 | 贡献率 | 累积贡献率 |
| 0.13161 | 0.017321 | 0.170306 | 0.655946 | 0.655946 |
| 0.069681 | 0.004855 | 0.04774 | 0.183872 | 0.839818 |
| 0.048169 | 0.00232 | 0.022814 | 0.087868 | 0.927686 |
| 0.035818 | 0.001283 | 0.012614 | 0.048585 | 0.976271 |
| 0.022939 | 0.000526 | 0.005174 | 0.019927 | 0.996198 |
| 0.01002 | 0.0001 | 0.000987 | 0.003802 | 1 |

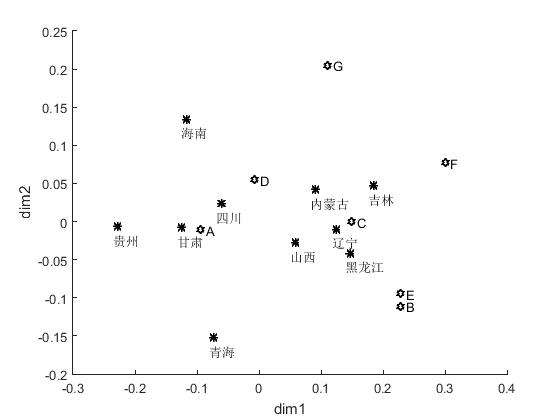


图1 行点和列点的散布图

总统计量等于0.2596，总统计量的83.98%可用前两维即可说明，它表示行点和列点之间的关系用二维表示就足够了。

在图1中，给出10个样品点（用1,2，…，10表示）和7个变量点（用A,B,…,G表示）在相同坐标系上绘制的散布图。从图中可以看出，样品点和变量点可以分为两类；第一类包括变量点B,C,E,F,G和样品点1,2,3,4,5；第二类包括变量点A,D和样品点6,7,8,9, 10。

在第一类中，变量为衣着（B），居住（C），医疗保健（E），交通和通讯（F）， 文教娱乐、日用品及服务（G）的支出分别占总支出的比重；地区有：山西（1），内蒙 古（2），辽宁（3），吉林（4），黑龙江（5），它们位于我国的东部和北部地区，说 明这5个地区的消费支出结构相似。在第二类中，变量为食品（A），家庭设备及服务（D） 的支出分别占总支出的比重；地区有：海南（6），四川（7），贵州（8），甘肃（9），青海（10）,它们位于我国的南部和西部地区，说明这5个地区的消费支出结构相似。

MATLAB源代码：

clc, clear

a=load('xf.txt'); %原始文件保存在纯文本文件xf.txt中

T=sum(sum(a));

P=a/T; %计算对应矩阵P

r=sum(P,2); c=sum(P); %计算边缘分布

Row\_prifile=a./repmat(sum(a,2),1,size(a,2)); %计算行轮廓分布阵

B=(P-r\*c)./sqrt((r\*c)); %计算标准化数据B

[u,s,v]= svd(B,'econ'); %对标准化后的数据阵B作奇异值分解

d=diag(s);

fprintf('奇异值为：\n');

disp(d);

w1=sign(repmat(sum(v),size(v,1),1)); %修改特征向量的符号矩阵，使得v中的每一个列向量的分量和大于0

w2=sign(repmat(sum(v),size(u,1),1)); %根据v对应地修改u的符号

vb=v.\*w1; %修改特征向量的正负号

ub=u.\*w2; %修改特征向量的正负号，本例中样本点个数和变量个数不等

lamda=diag(s).^2; %计算Z'\*Z的特征值，即计算主惯量

fprintf('主惯量为：\n');

disp(lamda);

ksi2square=T\*(lamda); %计算卡方统计量的分解

fprintf('卡方为：\n');

disp(ksi2square);

T\_ksi2square=sum(ksi2square); %计算总卡方统计量

fprintf('总卡方统计量为：\n');

disp(T\_ksi2square);

con\_rate=lamda/sum(lamda); %计算贡献率

cum\_rate=cumsum(con\_rate); %计算累积贡献率

fprintf('累积贡献率为：\n');

disp(cum\_rate);

beta=diag(r.^(-1/2))\*ub; %求加权特征向量

G=beta\*s; %求行轮廓坐标

alpha=diag(c.^(-1/2))\*vb; %求加权特征向量

F=alpha\*s; %求列轮廓坐标F

num1=size(G,1); %样本点的个数

rang=minmax(G(:,[1,2])'); %坐标的取值范围

delta=(rang(:,2)-rang(:,1))/(5\*num1); %画图的标注位置调整量

ch={'A', 'B', 'C', 'D', 'E', 'F', 'G'};

yb={'山西','内蒙古','辽宁','吉林','黑龙江','海南','四川','贵州','甘肃','青海'};

hold on

plot(G(:,1),G(:,2),'\*','Color','k','LineWidth',1.3) %画行点散布图

text(G(:,1)-delta(1),G(:,2)-3\*delta(2),yb) %对行点进行标注

plot(F(:,1),F(:,2),'H','Color','k','LineWidth',1.3) %画列点散布图

text(F(:,1)+delta(1),F(:,2),ch) %对列点进行标注

xlabel('dim1'), ylabel('dim2')

xlswrite('tt',[diag(s),lamda,ksi2square,con\_rate,cum\_rate])

%把计算结果输出到Excel文件，这样便于把数据直接贴到word中的表格

ind1=find(G(:,1)>0); %根据行坐标第一维进行分类

rowclass=yb(ind1); %提出第一类样本点

ind2=find(F(:,1)>0); %根据列坐标第一维进行分类

colclass=ch(ind2); %提出第一类变量

fprintf('第一类样本及变量：\n');

disp(rowclass);

fprintf('\n')

disp(colclass);

7偏最小二乘回归分析

偏小二乘回归提供一种多对多线性回归建模的方法，特别当两组变量的个数很多，且都存在多重相关性，而观测数据的数量（样本量）又较少时，用偏小二乘回归建立的模型具有传统的经典回归分析等方法所没有的优点。

偏小二乘回归分析在建模过程中集中了主成分分析，典型相关分析和线性回归分析方法的特点，因此在分析结果中，除了可以提供一个更为合理的回归模型外，还可以同时完成一些类似于主成分分析和典型相关分析的研究内容，提供更丰富、深入的一些信息。

例 本节采用兰纳胡德（Linnerud）给出的关于体能训练的数据进行偏小二乘回归建模。在这个数据系统中被测的样本点，是某健身俱乐部的20位中年男子。被测变量分为两组。第一组是身体特征指标X，包括：体重、腰围、脉搏。第二组变量是训练结果指标Y，包括：单杠、弯曲、跳高。原始数据见表1。

表2给出了这6个变量的简单相关系数矩阵。从相关系数矩阵可以看出，体重与腰围是正相关的；体重、腰围与脉搏负相关；而在单杠、弯曲与跳高之间是正相关的。从两组变量间的关系看，单杠、弯曲和跳高的训练成绩与体重、腰围负相关，与脉搏正相关。

表 1 体能训练数据

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No | 体重() | 腰围() | 脉搏() | 单杠() | 弯曲() | 跳高() |
| 1 | 191 | 36 | 50 | 5 | 162 | 60 |
| 2 | 189 | 37 | 52 | 2 | 110 | 60 |
| 3 | 193 | 38 | 58 | 12 | 101 | 101 |
| 4 | 162 | 35 | 62 | 12 | 105 | 37 |
| 5 | 189 | 35 | 46 | 13 | 155 | 58 |
| 6 | 182 | 36 | 56 | 4 | 101 | 42 |
| 7 | 211 | 38 | 56 | 8 | 101 | 38 |
| 8 | 167 | 34 | 60 | 6 | 125 | 40 |
| 9 | 176 | 31 | 74 | 15 | 200 | 40 |
| 10 | 154 | 33 | 56 | 17 | 251 | 250 |
| 11 | 169 | 34 | 50 | 17 | 120 | 38 |
| 12 | 166 | 33 | 52 | 13 | 210 | 115 |
| 13 | 154 | 34 | 64 | 14 | 215 | 105 |
| 14 | 247 | 46 | 50 | 1 | 50 | 50 |
| 15 | 193 | 36 | 46 | 6 | 70 | 31 |
| 16 | 202 | 37 | 62 | 12 | 210 | 120 |
| 17 | 176 | 37 | 54 | 4 | 60 | 25 |
| 18 | 157 | 32 | 52 | 11 | 230 | 80 |
| 19 | 156 | 33 | 54 | 15 | 225 | 73 |
| 20 | 138 | 33 | 68 | 2 | 110 | 43 |
| 均值  标准差 | 178.6  24.6905 | 35.4  3.202 | 56.1  7.2104 | 9.45  5.2863 | 145.55  62.5666 | 70.3  51.2775 |

表 2 相关系数矩阵

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  | 1 | 0.8702 | -0.3658 | -0.3897 | -0.4931 | -0.2263 |
|  | 0.8702 | 1 | -0.3529 | -0.5522 | -0.6456 | -0.1915 |
|  | -0.3658 | -0.3529 | 1 | 0.1506 | 0.225 | 0.0349 |
|  | -0.3897 | -0.5522 | 0.1506 | 1 | 0.6957 | 0.4958 |
|  | -0.4931 | -0.6456 | 0.225 | 0.6957 | 1 | 0.6692 |
|  | -0.2263 | -0.1915 | 0.0349 | 0.4958 | 0.6692 | 1 |

MATLAB源代码：

clc,clear

ab0=load('pz.txt'); %原始数据存放在纯文本文件pz.txt中

mu=mean(ab0);sig=std(ab0); %求均值和标准差

rr=corrcoef(ab0); %求相关系数矩阵

ab=zscore(ab0); %数据标准化

a=ab(:,[1:3]);b=ab(:,[4:end]); %提出标准化后的自变量和因变量数据

[XL,YL,XS,YS,BETA,PCTVAR,MSE,stats] =plsregress(a,b);

xw=a\XS; %求自变量提出成分系数，每列对应一个成分，这里xw等于stats.W

yw=b\YS; %求因变量提出成分的系数

a\_0=PCTVAR(1,:);b\_0=PCTVAR(2,:);

a\_1=cumsum(a\_0);b\_1=cumsum(b\_0);

i=1;%赋初始值

%判断提出成分对的个数

while ((a\_1(i)<0.9)&(a\_0(i)>0.05)&(b\_1(i)<0.9)&(b\_0(i)>0.05))

i=i+1;

end

ncomp=i;

fprintf('%d对成分分别为：\n',ncomp);

for i=1:ncomp

fprintf('第%d对成分：\n',i);

fprintf('u%d=',i);

for k=1:3%此处为变量x的个数

fprintf('+(%f\*x\_%d)',xw(k,i),k);

end

fprintf('\n');

fprintf('v%d=',i);

for k=1:3%此处为变量y的个数

fprintf('+(%f\*y\_%d)',yw(k,i),k);

end

fprintf('\n');

end

[XL2,YL2,XS2,YS2,BETA2,PCTVAR2,MSE2,stats2] =plsregress(a,b,ncomp);

n=size(a,2); m=size(b,2);%n是自变量的个数,m是因变量的个数

beta3(1,:)=mu(n+1:end)-mu(1:n)./sig(1:n)\*BETA2([2:end],:).\*sig(n+1:end); %原始数据回归方程的常数项

beta3([2:n+1],:)=(1./sig(1:n))'\*sig(n+1:end).\*BETA2([2:end],:); %计算原始变量x1,...,xn的系数，每一列是一个回归方程

fprintf('最后得出如下回归方程：\n')

for i=1:3%此处为变量y的个数

fprintf('y%d=%f',i,beta3(1,i));

for j=1:3%此处为变量x的个数

fprintf('+(%f\*x%d)',beta3(j+1,i),j);

end

fprintf('\n');

end

bar(BETA2','k') %画直方图

fprintf('因变量y的预测值：\n');

yhat=repmat(beta3(1,:),[size(a,1),1])+ab0(:,[1:n])\*beta3([2:end],:); %求y1,..,ym的预测值

disp(yhat);

ymax=max([yhat;ab0(:,[n+1:end])]); %求预测值和观测值的最大值

%下面画y1,y2,y3的预测图，并画直线y=x

figure, subplot(2,2,1)

plot(yhat(:,1),ab0(:,n+1),'\*',[0:ymax(1)],[0:ymax(1)],'Color','k')

legend('单杠成绩预测图',2)

subplot(2,2,2)

plot(yhat(:,2),ab0(:,n+2),'O',[0:ymax(2)],[0:ymax(2)],'Color','k')

legend('弯曲成绩预测图',2)

subplot(2,2,3)

plot(yhat(:,3),ab0(:,end),'H',[0:ymax(3)],[0:ymax(3)],'Color','k')

legend('跳高成绩预测图',2)

该MATLAB源代码结果分析：根据PCTVAR可知，前两个成分解释自变量的比率为92.13%，只需要取两对成分即可。故ncomp的值应为2，beta3为最后求解出的回归方程。

，

，

。

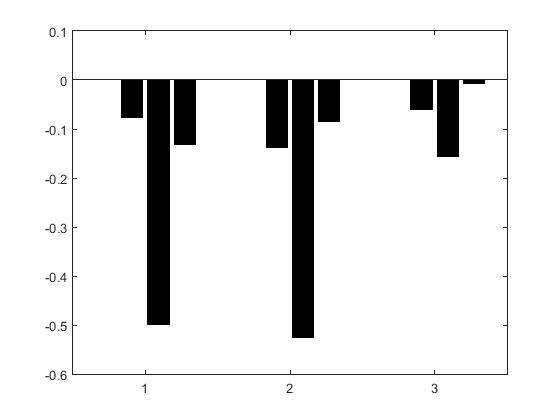


图1 回归系数的直方图

