一、聚类分析【系统聚类法、动态聚类法】

Q 型分析是对**样品**进行分类，**R 型**分析是对**变量**进行分类。

1.系统聚类法

```
proc cluster data=development method=single outtree=tree standard;
 id region;
 var x1-x10;
 run;
 proc tree data=tree horizontal graphics;
 id region;
 run;
 proc cluster data=development method=average outtree=tree standard;
 id region;
 var x1-x10;
 run;
 proc tree data=tree horizontal graphics;
 id region;
 run;
 proc cluster data=development method=centroid outtree=tree standard;
 id region;
 var x1-x10;
 run;
 proc tree data=tree horizontal graphics;
 id region;
 run;
 proc cluster data=development method=ward outtree=tree standard;
 id region;
 var x1-x10;
 run;
 proc tree data=tree horizontal graphics;
 id region;
 run;
```

“proc cluster”是一个系统聚类过程，“data=”指定所要分析的数据集。“method=”指定系统聚类方法，包括 single 最短距离法、complete 最长距离法、median 中间距离法、centroid 重心法、average 类平均法和 ward 离差平方和法。“outtree=”将聚类的过程输出到一个数据集，根据这个数据绘制聚类谱系图。“standard”选项将原始数据标准化为均值为 0、标准差为 1 的数据。“id region”指定以 region 作为各个样品的标签。“var x1-x10”指定变量为 x1-x10，缺省时默认为全部定量变量。“run”指示程序运行。

“proc tree”是一个画聚类谱系图的过程，“data=”指定用来画聚类谱系图的数据集，缺省时默认为 cluster 过程所产生的数据集。“horizontal”选项指定输出水平的树状图，“graphics”选项指示使用高分辨率图形。“id region”指定聚类谱系图中以 region 的变量值作为标签，“run”指示程序运行。

1.类分析统计量，包括特征值，解释变异的比重和累计比重：

相关矩阵的特征值

	特征值	差分	比例	累积
1	5.33897208	3.44631986	0.5339	0.5339
2	1.89265222	0.71170234	0.1893	0.7232
3	1.18094987	0.47584415	0.1181	0.8413

2.聚类历史：

聚类历史

聚类数	连接聚类		频数	Norm Minimum Distance	结值
10	湖州市	绍兴市	2	0.4186	
9	金华市	台州市	2	0.4262	
8	温州市	CL9	3	0.4348	

最短距离法：

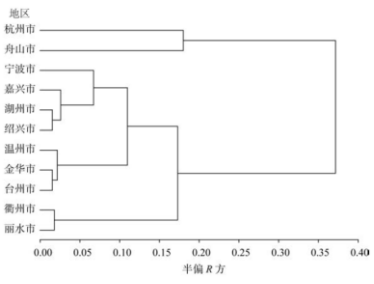
输出 3-3 为最短距离法的聚类过程，“Norm Minimum Distance(正规化最短距离)”等于合并的两类之间的距离除以样品间的平均距离。

聚类历史

聚类数	连接聚类		频数	半偏 R 方	R 方	结值
10	湖州市	绍兴市	2	0.0151	.985	
9	金华市	台州市	2	0.0156	.969	
8	衢州市	丽水市	2	0.0183	.951	

离差平方和法：

输出 3-2 为离差平方和法的聚类过程，“聚类数”是新类形成后类的个数，“连接聚类”是一个聚类过程中相连接的两类，“频数”是新类中包含的样品个数。例如：CL9 由金华市和台州市连接而成，在下一个聚类过程中与温州市连接成为 CL7，此时该类中有三个样品。“半偏 R 方”与“R 方”这两个统计量是用来帮助确定分类个数的。R 方越大，表示类之间区分得越开，聚类效果越好。然而，不能简单地以 R 方的大小确定分类个数，还应考察其值的变化，也就是半偏 R 方。半偏 R 方较大，说明本次并类的效果不好，应当考虑聚类到上一步停止。在本例中，半偏 R 方最大为聚为一类时，也就是说，聚为两类是比较合理的。然而，聚类的类数还应结合树状图、经济含义等综合得出。



3.树形图：人为给定阈值以确定聚类类数。如给定阈值 0.15，11 个地区聚为 4 类。

4.分析结果。某类包括谁谁谁 xxx，结合经济意义说聚为几类更好，然后具体的类别情况是怎么样的。

2.动态聚类法

```
proc standard data=lifeq m=0 std=1 out=sv;
var x1-x11;
run;
data a1;
set sv;
if _n_=10 then output;
if _n_=18 then output;
if _n_=27 then output;
run;
proc fastclus maxclusters=3 data=sv seed=a1 mean=stat out=out_data;
var x1-x11;
run;
```

if _n_=10 或 18 或 27，根据题意判断是否需要进行更改，如果有说就改成对应的凝聚点。凝聚点的选择好坏直接影响聚类结果。

“proc standard”是对数据进行标准化的过程，“data=”指定原始数据存放的数据集，“m=”和“std=”分别指定了标准化数据的均值和标准差，“out=”指定输出数据存放的数据集。与系统聚类不同，使用 SAS 软件对原始数据进行动态聚类时无法直接使用 standard 选项。因此，在动态聚类前，必须自行对数据进行标准化。

“data a1”创建一个新数据集保存动态聚类的凝聚点数据，未指定逻辑库时数据集默认保存在“work”逻辑库中。“set sv”调用了标准化处理后的相关数据，“if _n_=”指定在 a1 中要保留的观测(本例中保留第 10、18、27 个观测，即江苏、湖南、甘肃)。本例使用主观判断法选取凝聚点，读者也可自行使用重心法、均值法等其他方法进行选取。

“procfastclus”是一个动态聚类过程，“maxclusters=”指定所允许的最大分类个数。“data=”指定要进行动态聚类的数据集。“seed =”指定一个 SAS 数据集，其中包括要选择的初始凝聚点。“mean =”生成一个输出数据集，其中包含每个类的均值和一些统计量。“out =”生成一个输出数据集，其中包含原始数据和新变量聚类以及与聚类种子的距离。var 语句指定参与聚类过程的变量。

聚类汇总

聚类	频数	均方根标准差	从种子到观测的最大距离	半径超出	最近的聚类	聚类质心间的距离
1	6	1.1434	6.3846		2	5.2254
2	15	0.5289	3.6915		3	2.1825
3	9	0.4350	2.4669		2	2.1825

1.聚类汇总：说明动态聚类完成后每一类的频数、标准差以及与凝聚点的最大距离等信息。

2.分析结果。在 out_data 中看具体的聚类情况，然后弄成表格进行展示、对各个类进行分别命名和分析。

表 3-13　全国各地区居民生活质量聚类

类　别	地　区
高生活质量地区	北京、天津、上海、江苏、浙江、广东
中等生活质量地区	河北、内蒙古、辽宁、安徽、福建、江西、山东、湖北、湖南、重庆、四川、陕西、宁夏、新疆
低生活质量地区	山西、吉林、黑龙江、河南、广西、海南、贵州、云南、甘肃、青海

二、判别分析【距离判别法、费歇判别法】

1.距离判别法

```
Proc discrim data=classified list listerr testdata=unclassified
out=classified_out testout=unclassified_out outstat=os pool=yes method=normal;
Class type;
Var X1-X11;
Run;
```

这里 class type 的 type 要改成对应训练样本的已分类标签。

“Proc discrim”是一个距离判别过程，选项“list”是指输出所有训练样本的判别结果，选项“listerr”是指列出训练样本中所有误判的结果，选项“data=classified”是指已分类的数据集是“classified”，选项“testdata=unclassified”是指要分类的数据集是“unclassified”，选项“out=classified_out”是指已分类数据的回判结果生成为“classified_out”，选项“testout=unclassified_out”是指要分类数据的判别结果生成为“unclassified_out”，选项“outstat=os”是指计算过程中的一些统计量生成为“os”数据集。

此外，选项“method=normal”是指各总体均服从于多元正态分布，可缺省。选项“pool=yes”是指采用合并的协方差矩阵，当各总体协方差矩阵相等时得出线性判别函数；如果各总体协方差矩阵不相等时，则使用选项“pool=no”可得出二次判别函数；默认是选项“pool=yes”。

在考虑先验概率时，需使用 priors 语句，该语句缺省时为“priors equal”，表示默认各组的先验概率相等。如果希望各组先验概率与训练样本中各组所占比例相等，可增加语句“priors proportional”，亦可增加语句来指定先验概率。例如：priors ‘ST’=0.1 ‘非 ST’=0.9。

Class 语句指定分组变量。Var 语句是指定要分析的变量。

1.展示这个表，并写出具体的判别函数和判别规则。

以下对象的线性判别函数: TYPE

变量	标签	ST	非ST
常数		-2.92421	-2.07285
X1	总资产收益率	-6.80562	1.51080
X2	净资产收益率	0.03062	-0.04370
X3	营业净利率	-0.56583	-0.08834

第一类判别函数为：

$$f_1(x) = -2.924\,2 - 6.805\,6x_1 + 0.030\,6x_2 - 0.565\,8x_3 + 0.013\,9x_4 + 0.564\,1x_5 + 0.597\,6x_6 - 1.258\,8x_7 - 4.337\,3x_8 + 0.199\,1x_9 - 0.086\,8x_{10} - 0.142\,2x_{11}$$

第二类判别函数为

$$f_2(x) = -2.072\,9 + 1.510\,8x_1 - 0.043\,7x_2 - 0.088\,3x_3 + 0.004\,8x_4 + 3.057x_5 - 0.109\,4x_6 - 0.606x_7 + 7.925\,9x_8 + 0.423\,7x_9 - 0.233x_{10} + 0.176\,5x_{11}$$

哪一个判别函数的取值大，就归为哪一类。

2.展示训练样本的回代结果正确率，并给出文字说明：

以下校准数据的分类汇总: AA.ENTERPRISE_CLASSIFIED
使用以下项的重新替换汇总: 线性判别函数

分入“TYPE”的观测数和百分比

从 TYPE	ST	非ST	合计
ST	16 80.00	4 20.00	20 100.00
非ST	1 3.33	29 96.67	30 100.00
合计	17 34.00	33 66.00	50 100.00
先验	0.5	0.5	

“TYPE”的出错数估计

	ST	非ST	合计
比率	0.2000	0.0333	0.1167
先验	0.5000	0.5000	

输出 4-3 表明：在训练样本中，20 家 ST 企业中有 16 家企业判断为 ST 类别，4 家企业判断为非 ST 类别，错判率为 20%；30 家非 ST 企业中有 1 家企业判断为 ST 类别，29 家企业判断为非 ST 类别，错判率为3.33%；合计的错判率为 11.67%。

如果有需要，可以找出具体的错误判别样本：

成员的后验概率TYPE

观测	从 TYPE	分为TYPE		ST	非ST
1	非ST	非ST		0.1516	0.8484
2	非ST	非ST		0.0635	0.9365
3	非ST	非ST		0.0176	0.9824
4	非ST	非ST		0.3596	0.6404
5	非ST	非ST		0.2351	0.7649
6	非ST	非ST		0.0297	0.9703
7	非ST	ST	*	0.6538	0.3462

带*的就是错误判别的。

分入“TYPE”的观测数和百分比

	ST	非ST	合计
合计	9 18.75	39 81.25	48 100.00
先验	0.5	0.5	

3.展示测试样本判别结果：说明一下结果，在 unclassified_out 中找到具体情况并表格展示。

2.费歇判别法（a 就是 A）

```
Proc candisc data=flower out=flower_out;
class a;
var x1-x4;
run;
Proc plot data=flower_out;
plot can2*can1=A;
run;
```

记得要把 class a 中的 a 和 can2*can1=A 中的 A 换成对应数据集的类别标签。

“proc candisc”是一个典型判别过程，输出数据集选项“out=flower_out”要求生成一个包含原始数据和典型变量得分即 can1 和 can2 的 SAS 数据集，并命名为 flower_out。

“proc plot”是一个图形过程。语句“plot can2* can1=type”要求作散点图，can2 为垂直变量，can1 为水平变量，用变量 type 的值作为散点的标记。

1.给出特征值情况(框起来的部分)，确定判别函数个数。

	典型相关	调整典型相关	近似标准误差	典型相关平方	特征值: Inv(E)*H = CanRsq/(1-CanRsq) 特征值	差分	比例	累积	H0 检验: 当前行和之后的所有行的典型相关都是零 似然比	近似 F 值	分子自由度	分母自由度	Pr > F
1	0.984821	0.984508	0.002468	0.969872	32.1919	31.9065	0.9912	0.9912	0.02343863	199.15	8	288	<.0001
2	0.471197	0.461445	0.063734	0.222022	0.2854		0.0088	1.0000	0.77797337	13.79	3	145	<.0001

费歇判别中，最多可以得到 $\min(k-1, p)$ 个判别函数，其中 k 为类别数，p 为变量个数。因此，在本案例中，共可以得到两个判别函数。

由输出 4-5 可知，矩阵 $\boldsymbol{E}^{-1}\boldsymbol{B}$ 最大的特征值为32.191 9，它的方差贡献率为99.12%，即第一判别函数解释了99.12% 的方差，第二判别函数解释了0.88% 的方差，这两个判别函数能解释全部的方差。根据判别函数个数的选取规则，选取一个判别函数就可以了。

2.给出中心化的费歇判别函数：

原始典型系数

变量	标签	Can1	Can2
X1	花萼长	-.0829377642	0.0024102149
X2	花萼宽	-.1534473068	0.2164521235
X3	花瓣长	0.2201211656	-.0931921210
X4	花瓣宽	0.2810460309	0.2839187853

均值在 os 里面。

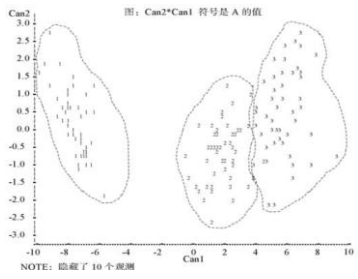
由输出 4-6 可以得到中心化的费歇判别函数为

$$y_1 = -0.082\,9(x_1 - 58.433) + 0.152\,4(x_2 - 30.573) + 0.220\,1(x_3 - 37.580) + 0.281\,0(x_4 - 11.993)$$

$$y_2 = 0.002\,4(x_1 - 58.433) + 0.216\,5(x_2 - 30.573) - 0.093\,2(x_3 - 37.580) + 0.283\,9(x_4 - 11.993)$$

根据该判别函数可以计算出每个观测的判别函数得分(y_1, y_2)。

3.判别的散点图：



类别的虚线框自己截图的时候加一下，看得更直观。

输出 4-8 是将 150 个样品的判别函数得分(y_1, y_2)做散点图得到的结果，图中 can1 和 can2 分别指 y_1, y_2，“1”代表 setosa 鸢尾花，“2”代表 versico-lor 鸢尾花，“3”代表 virginica 鸢尾花。图中显示有 10 个隐藏点，这是由于这些样本点与图中的某个样本点几乎重叠，因而未能被标出。

从输出 4-8 中可以看出，三组分离的效果非常好，且分离的很大程度显示在 can1 上，这与一个判别函数解释的方差贡献率相符合。因此，对于一个新的待判样本，通过计算判别得分，在图中标出坐标点，就可以判断出新样本点的所属类型。

三、主成分分析

【主成分回归时，主成分的含义不需要解释（只是为了避免多重共线性）。其他情形下，均要解释主成分含义。】

R 型分析是利用相关系数矩阵来进行分析，Q 型分析是利用协方差矩阵来进行分析。一般 R 型分析效果更好。

1.主成分分析

```
proc princomp data=economic out=prin;
run;
ods graphics on;
proc princomp data=economic out=prin n=3 plot=pattern(ncomp=2) plot=
score(ncomp=2);
var x1-x10;
id region;
run;
ods graphics off;
proc plot data=prin;
plot prin2*prin1 $ region='*'/href=0 vref=0;
run;
```

“proc princomp”是一个进行主成分分析的过程。选项“data=econonmic”是指定使用数据集“economic”进行主成分分析；选项“out=prin”是要求生产一个包含原始数据及主成分得分的 SAS 数据集，并取名为 prin，逻辑库库名缺省是指保存在临时库 work；选项“n=3”是指定输出的主成分为三个，当其缺省时，表示输出所有的主成分；选项“plot=pattern(ncomp=2)”是指建立以$\rho(prin_1, x_i)$为横坐标、$\rho(prin_2, x_i)$为纵坐标的散点图，在坐标系中描出各变量对应的点(图解变量)；选项“plot=score(ncomp=2)”是指建立以 $prin_1$ 为横轴、$prin_2$ 为纵轴，以 $prin_1$ 得分为横坐标、$prin_2$ 得分为纵坐标的散点图，在坐标系中描绘出每个省份所对应的点(图解样品)。

如果基于协方差矩阵出发做主成分分析，则需要加上选项“cov”，该选项缺省时是从相关系数矩阵出发进行分析。此外，该过程步中还省略了选项“prefix”。若选项“prefix=F”，则表示在输出数据集 prin 中，主成分变量是 F_1, F_2, …。读者可根据研究需要进行添加和更改。

“proc plot”是绘制散点图的过程步，在此案例中是为了单独绘制图解样品的图。

1.展示相关矩阵：

相关矩阵

		x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
x1	人均地区生产总值	1.0000	0.6797	0.8953	0.9324	0.8920	0.6742	0.9123	0.8087	0.6655	0.6587
x2	第三产业产值所占比重	0.6797	1.0000	0.7599	0.7891	0.8298	0.5998	0.8104	0.6238	0.7212	0.6668
x3	人均进出口额	0.8953	0.7599	1.0000	0.9617	0.9614	0.7121	0.9005	0.8023	0.5875	0.7139

在计算各变量的特征值之前，需对各变量之间的相关性进行分析，输出 5-1 是各变量之间的相关系数矩阵。在相关系数矩阵中可以看到各变量之间的相关系数较高，说明有必要通过主成分分析对变量进行降维处理。

2.展示相关矩阵的特征值：

相关矩阵的特征值

	特征值	差分	比例	累积
1	7.84106585	7.24417795	0.7841	0.7841
2	0.59688790	0.12197055	0.0597	0.8438
3	0.47491735	0.07670241	0.0475	0.8913
4	0.39821494	0.06627548	0.0398	0.9311

输出 5-2 显示了原始数据相关系数矩阵特征值(主成分的方差)、方差贡献率以及累计方差贡献率。可以看出第一个特征值为7.84，第一个主成分的方差贡献率为78.41%，前两个主成分的累计方差贡献率达84.38%，可以解释大部分的变量信息。因此在本案例中，选取两个主成分进行分析。

3.展示特征向量，写出主成分表达式：

特征向量

		Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
x1	人均地区生产总值	0.330116	-.213293	-.279443	-.013590	0.035970	-.474181	0.605267	0.344369	0.023692	0.235829
x2	第三产业产值所占比重	0.302123	0.326245	0.235686	0.044103	-.679453	0.388817	0.323098	0.110407	0.118335	0.037042
x3	人均进出口额	0.337460	-.272227	-.088823	-.258711	-.159637	0.033447	-.429651	-.205547	0.304983	0.625081
x4	人均可支配收入	0.345657	-.136500	-.171382	-.033306	-.164891	-.157735	-.377078	0.321851	0.349186	-.641586

输出 5-3 中的表格是前三个主成分的特征向量。表中 $prin_1$、$prin_2$ 分别表示第一、二主成分，根据特征向量可以写出前两个主成分的表达式：

$$prin_1 = 0.330\ 1x_1 + 0.302\ 1x_2 + 0.337\ 5x_3 + 0.345\ 7x_4 + 0.346\ 0x_5 + 0.276\ 8x_6 + 0.346\ 5x_7 + 0.301\ 0x_8 + 0.279\ 0x_9 + 0.285\ 5x_{10}$$

$$prin_2 = -0.213\ 3x_1 + 0.336\ 2x_2 - 0.272\ 2x_3 - 0.136\ 5x_4 - 0.075\ 9x_5 - 0.415\ 1x_6 - 0.047\ 5x_7 - 0.067\ 7x_8 + 0.660\ 1x_9 + 0.367\ 0x_{10}$$

4.分析结果。解释主成分含义。主成分得分在 prin 里。

表 5-12　主成分得分

地区	第一主成分得分	第二主成分得分	第一主成分得分排名
北京	9.975 9	1.464 3	1
上海	7.417 3	−0.938 7	2
天津	3.568 3	−0.566 6	3
浙江	2.486 2	0.353 7	4

由表 5-12，按第一主成分 $prin_1$ 的得分排序，综合发展水平处于全国平均水平之上的依次为：北京、上海、天津、浙江、广东、江苏、福建、山东、辽宁。综合发展水平一般的地区有：内蒙古、重庆、宁夏、海南、湖北、新疆、青海、陕西、吉林、安徽、山西、湖南、黑龙江。综合发展水平较落后的地区有：河北、四川、贵州、甘肃、江西、广西、河南和云南。

如果有需要，还可以再加图解样本和图解变量的散点图进行分析，说明样本或变量的分类情况。

2.主成分回归

①分步构建主成分回归模型

```
proc corr data=princomp_reg;
var y x1-x3;
run;
proc reg data=princomp_reg;
model y=x1 x2 x3;
run;
proc standard data=princomp_reg out=sv mean=0 std=1;
var y x1-x3;
run;
proc princomp data=sv out=opcr;
var x1-x3;
run;
proc reg data=opcr ;
model y=prin1 prin2;
run;
quit;
```

“proc corr”是进行相关分析的过程步，“var y x1-x3”是指求变量 $X_1 - X_3$ 以及变量 Y 之间的相关系数。

“proc standard”是进行标准化变换的过程步，对变量 Y，$X_1 - X_3$ 进行标准化变换。

“proc princomp”是主成分分析的过程步，选项“data=sv”是指对 sv 数据集进行主成分分析，选项“out=opcr”是指生成一个数据集 opcr，保存原始数据和主成分得分。

“proc reg”是进行回归分析的过程步。语句“model y=prin1 prin2”是指以 Y 为因变量，以 $prin_1$，$prin_2$ 作为自变量构建回归模型。

Pearson 相关系数，N = 31
Prob > |r| under H0: Rho=0

	y	x1	x2	x3
y 人均消费支出	1.00000	0.96199 <.0001	0.97077 <.0001	0.81109 <.0001
x1 人均生产总值	0.96199 <.0001	1.00000	0.93280 <.0001	0.81277 <.0001

1.相关性分析：

在进行回归分析之前，需要先计算出各个自变量和因变量之间的相关系数，结果如输出 5-7 所示，可以看到 X_1 与 Y 的相关系数为0.962 0，且 $p < 0.000\ 1$，表明 X_1 与 Y 存在高度线性相关；同样地，自变量 X_2、X_3 与 Y 也存在高度线性相关。在此基础上进一步做回归分析。

2.回归分析：展示回归模型的显著性结果，说明 F 统计量。

方差分析

源	自由度	平方和	均方	F 值	Pr > F
模型	3	3222734927	1074244976	267.20	<.0001
误差	27	108550866	4020402		
校正合计	30	3331285794			

展示 R^2 的那个表，说明回归方程拟合程度。

3.模型检验：展示参数估计结果。

参数估计

变量	标签	自由度	参数估计	标准误差	t 值	Pr > |t|
Intercept	Intercept	1	259.78566	2755.96615	0.09	0.9256
x1	人均生产总值	1	0.17056	0.03750	4.55	0.0001
x2	人均可支配收入	1	0.58614	0.10249	5.72	<.0001

输出 5-10 给出了回归模型的参数估计结果。X_1 和 X_2 的参数估计值为0.170 6、0.586 1，T 检验统计量的值较大，p 值为0.000 1，在0.01 的显著性水平下通过检验。但 X_3 的参数估计值为 −46.096 8，T 检验统计量的值较小，这与前面相关分析的结果不一致，不符合实际情况。

各变量之间可能存在多重共线性导致回归估计结果失真，因此考虑采用主成分回归分析方法。

相关矩阵的特征值

	特征值	差分	比例	累积
1	2.71887483	2.50414611	0.9063	0.9063
2	0.21472872	0.14833226	0.0716	0.9779

4.主成分分析。展示特征值情况，选择主成分个数。

输出 5-11 是对 3 个自变量进行主成分分析时，相关系数矩阵的特征值、方差贡献以及累积方差贡献率。第一主成分的特征值为2.718 9，方差贡献率为90.63%，前两个主成分累积方差贡献率达到97.79%，大于 95%，因此根据主成分回归模型中选取主成分个数的规则，在本案例中选取 2 个主成分进行建模和分析。

根据特征向量的表，写出主成分表达式：

特征向量

		Prin1	Prin2	Prin3
x1	人均生产总值	0.583807	-.446401	0.678156
x2	人均可支配收入	0.587671	-.343982	-.732338
x3	互联网普及率	0.560190	0.826077	0.061518

输出 5-12 是各个主成分的特征向量，根据特征向量可以写出前两个主成分的表达式：

$$prin_1 = 0.583\ 8X_1 + 0.587\ 7X_2 + 0.560\ 2X_3 \quad (5.4)$$

$$prin_2 = -0.446\ 4X_1 - 0.344\ 0X_2 + 0.826\ 1X_3 \quad (5.5)$$

以居民消费水平为因变量，前两个主成分作为自变量，构建主成分回归模型进一步

5.构建主成分回归模型和检验。展示主成分回归模型的显著性结果，说明 F 统计量、调整 R^2。给出参数估计结果。

参数估计

变量	标签	自由度	参数估计	标准误差	t 值	Pr > |t|
Intercept	Intercept	1	1.978E-16	0.03414	0.00	1.0000
Prin1		1	0.58351	0.02104	27.73	<.0001
Prin2		1	-0.43466	0.07488	-5.80	<.0001

写出主成分回归模型。

输出 5-15 是主成分回归模型的参数估计结果，第一、第二主成分的 p 值都小于 0.000 1，在1%的显著性水平下通过了 t 检验。根据表格中参数估计值，可得到主成分回归模型

$$Y = 0.583\ 5prin_1 - 0.434\ 7prin_2 \quad (5.6)$$

②直接构建主成分回归模型【结果同①】

```
proc reg data=princomp_reg outest=out;
model y=x1-x3/pcomit=1, 2;
run;
quit;
proc print data=out;
run;
```

“proc reg”是回归分析的过程步，但是在语句“model y = x1 - x3”后加上选项“pcomit=1, 2”，表示剔除 1 个主成分和剔除 2 个主成分进行主成分回归。

“proc print”是输出数据结果的过程步。

四、因子分析

```
proc factor data=lifeq method=prin n=4 r=v out=out outstat=stat reorder;
var X1-X11;
run;
proc plot data=out;
plot factor2*factor1 $ region='*'/href=0 vref=0;
run;
data a1;
set out;
f=0.3863*factor1+0.3287*factor2+0.1144*factor3+0.0999*factor4;
keep region f factor1 factor2 factor3 factor4;
run;
proc sort data=a1;
by descending f;
run;
```

f=0.3863*factor1.系数是旋转后方差贡献除以总变量个数。

“proc factor”是一个进行因子分析的过程，“data=”指定要进行因子分析的原始数据集，“reorder”选项能使输出的因子载荷矩阵按载荷值大小重新排序。

“method=prin”指定使用主成分法进行因子分析，常见的还可使用“ML”(极大似然估计法)，“method”缺省时指的是主成分法；“n=”指定要保留的公因子个数，缺省状态下仅保留特征值大于 1 的因子；“r=v”是“rotate=varimax”的缩写，指定使用方差最大法正交旋转，常用的还有“r=q”(rotate=quartimax，四次方最大法)和“r=e”(rotate=equamax，等量最大法)，也可使用“PROMAX”(一种斜交旋转法)；“out=”指定因子分析的输出数据集，该数据集包含原始数据和公共因子得分；“outstat=”指定输出数据集中保存因子分析过程中的统计量。var 语句指定参与因子分析的变量，缺省时默认使用全部定量变量。

“proc plot”是一个作图的过程，“data=”指定用于作图的数据集。“plot factor2* factor1 $ region='* '/href=0 vref=0”语句绘制了一个以 factor1 为横轴、factor2 为纵轴，每一标签变量(此例中为 region)以“*”表示，水平参考线和垂直参考线经过(0，0)点的散点图。

“data a1”创建了一个名为“a1”的新数据集，未指定逻辑库时数据集默认保存在 work 库中。set 语句是将一个数据集中的数据复制到创建的新数据集中，本例通过 set 语句将 out 数据集中的数据复制到 a1 数据集中。“f=”是一个赋值语句，等式左边的“f”是变量名，等式右边是计算公式，本例以各公共因子的方差贡献率为权数计算综合得分。keep 语句用于指定数据集中要保留的变量。

“proc sort”是一个排序的过程，“data = a1”指定需要排序的数据集，“by descending f”指定对变量“f”进行降序排列。

1.给出初始的因子分析相关统计量。帮助确定因子个数。

下，SAS 系统自动保留特征值大于 1 的公共因子，而在实证分析中习惯取累积方差贡献率为 85% 这一阈值作为确定标准。前三个公共因子的累积方差贡献率为86.84%，表明能反映原始变量86.84% 的信息。

先验公因子方差估计: ONE

相关矩阵的特征值: 总计 = 11 平均值 = 1

	特征值	差分	比例	累积
1	7.51187581	6.41894506	0.6829	0.6829
2	1.09293075	0.14573680	0.0994	0.7823
3	0.94719395	0.27629257	0.0861	0.8684

比如确定因子个数为 4。

给出初始的公共因子载荷，并说明。

输出 6-2 为初始公共因子的因子载荷，它们是变量与公共因子的相关系数，绝对值越大，相关的密切程度越高。例如：X_2 在 $Factor_1$ 上的因子载荷为0.976 46，即 X_2 和 $Factor_1$ 的相关系数为0.976 46，两者高度相关。

因子模式

		Factor1	Factor2	Factor3	Factor4
x2	人均可支配收入	0.97646	0.10110	-0.09959	0.06811
x1	人均地区生产总值	0.94888	0.02208	-0.03292	-0.06519
x6	每百户计算机拥有量	0.83332	0.10445	-0.26579	0.09966

2.说明因子含义。给出旋转后的因子载荷矩阵并命名(先说明旋转了更好)。（数据为 **SAS 输出的旋转因子模型**）

输出 6-6 为因子旋转后的因子载荷矩阵，可明显看到，旋转后的各因子载荷值进一步拉大或缩小，使得公共因子较旋转前含义更加明确。

表 6-6　旋转后的因子载荷及因子命名

因子名	变量名		$Factor_1$	$Factor_2$	$Factor_3$	$Factor_4$
$Factor_1$：经济条件和精神生活因子	X_1	人均地区生产总值	**0.691 5**	0.605 2	0.192 0	−0.157 9
	X_2	人均可支配收入	**0.792 6**	0.512 1	0.277 7	−0.103 4
	X_3	人均拥有公共图书馆藏量	**0.940 8**	0.192 4	−0.007 9	−0.034 4
	X_4	人均教育文化娱乐消费支出	**0.867 7**	0.365 9	0.227 9	0.049 0
	X_5	人均交通通讯消费支出	**0.644 2**	0.687 4	0.101 5	−0.072 4
$Factor_2$：交通出行和物质生活	X_6	每百户计算机拥有量	0.842 0	**0.368 6**	0.285 8	−0.189 6
	X_7	每百户家用汽车拥有量	0.402 9	**0.749 2**	−0.006 7	−0.200 1

3.阐述公共因子得分。**数据在 a1。**

表 6-7　因子得分

名次	地区	经济条件和精神生活因子	地区	交通出行和物质生活因子	地区	生活环境因子	地区	医疗资源因子
1	上海	4.448 4	北京	3.123 5	北京	2.494 4	新疆	1.502 2
2	浙江	1.119 8	天津	2.079 6	江西	1.530 0	辽宁	1.474 9

4.计算综合得分。根据这个表确定因子综合得分的系数(除变量个数)。**数据在 a1。**

每个因子已解释方差

Factor1	Factor2	Factor3	Factor4
4.2491614	3.6162322	1.2582282	1.0992802

表 6-8　各地区居民生活质量综合得分

排名	地区	得分	排名	地区	得分
1	北京	1.561 1	16	四川	−0.089 6
2	上海	1.502 0	17	青海	−0.105 1
3	浙江	0.692 7	18	湖北	−0.143 3

五、对应分析【tice 是自己 cards 的数据】

```
proc corresp data=tice  out=tice_out rp cp short;
var ps zc cz fp;
label ps="偏瘦" zc="正常" cz="超重" fp="肥胖";
id level;
run;
```

“data tice”创建一个名为“tice”的新数据集，其中包含 level、ps、zc、cz、fp 五个变量。由于 level 是定性变量，需要在变量后加符号“$”进行识别。

“proc corresp”是一个进行对应分析的过程，选项“rp”和“cp”分别输出行轮廓矩阵和列轮廓矩阵，选项“short”是指不输出除坐标轴以外的所有点和坐标统计量。

var 语句指示列联表中的列是 ps、zc、cz、fp，label 对变量的标签进行设置，例如 ps 的标签为“偏瘦”，id 语句指定用 level 的值作为输出列联表中行的名称。

若体测数据是每个学生体质和成绩的原始数据而不是列联表形式，需要使用以下命令：

```
proc corresp data=tice out=results rp cp all;
tables row, column;
weight f;
run;
```

其中，tables 语句指定用于构造列联表的行变量和列变量，该语句的第一个变量规定为行变量，逗号后的第二个变量为列变量；在使用 tables 语句时不能使用 id 语句；weight 语句用来读入类别组合的频数。

在 SAS 9.4 版本中，运行对应分析命令后，软件会自行给出对应分析相关图。若想绘制个性化的图，可以增加以下过程步：

```
proc plot data=incmoe_out;
plot dim2*dim1="*"$  region /box vspace=6 hspace=15 haxis=-0.5 to 0.5 by .15 vaxis
=-0.5 to 0.5 by .15;
run;
```

“proc plot”是一个作图的过程。“dim2*dim1”表示横坐标是 dim2、纵坐标是 dim1，表中各点的位置用“*”表示，box 要求画出的边框围住整个图形。“vspace=6 hspace=15”规定了图中纵坐标、横坐标单位格在图中的实际长度。“haxis=-0.5 to 0.5 by .15”是指横坐标的刻度范围是−0.5至0.5，且单位刻度为0.15。

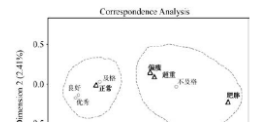
1.各维汇总表。说明主惯量解释列联表数据的变异的程度。

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	0 20 40 60 80 100
0.35244	0.12421	894.345	97.59	97.59	
0.05540	0.00307	22.099	2.41	100.00	

输出 7-3 是各维汇总表，其中 Singular 是奇异值，Principal inertia 是主惯量，Percent 是惯量的百分比，最后一列数据是惯量占比的累计值。从中可以看出，第一维和第二维的惯量比例占总惯量的 100%，因此前两维解释了列联表数据 100% 的变异。

2.对应分析图。



从图 7-4 中可以看到，体测成绩不及格与体质状况偏瘦、超重的距离较近，体测成绩及格与体质状况正常的距离较近。结果表明，体质状况与体测成绩有一定联系，体质状况正常与体测成绩及格关系密切。

六、典型相关分析

```
Proc cancorr  data=tech  out=techout outstat=techvalue all;
With Y1-Y5;
var X1-X4;
Run;
```

“Proc cancorr”是一个典型相关分析的过程。选项“data=tech”是指定分析的数据集；“out=techout”是指生成“techout”这个数据集，包括原始数据和典型变量得分；选项“outstat=techvalue”表示将分析得到的各种统计量生成到数据集“techvalue”中。此外，选项“all”是指对变量进行冗余分析，缺省时则不做冗余分析。

在该语句中缺省了“vprefix”和“wprefix”，缺省时表示第一组变量的前缀是 v，第二组变量的前缀是 w。当“vprefix= u”时表示第一组典型变量的前缀是 u，当“wprefix= v”时表示第二组典型变量的前缀是 v。

“With Y1 - Y5”指定分析的第二组变量为 Y_1 至 Y_5。

“var X1 - X4”指定分析的第一组变量为 X_1 至 X_4。

VAR 变量 和 WITH 变量 之间的相关性

	y1	y2	y3	y4	y5
x1	0.7329	0.6822	0.7216	0.6036	0.8366
x2	0.7631	0.7210	0.6503	0.5498	0.8868
x3	0.7339	0.5815	0.4740	0.0839	0.6488

1.给出 var 和 with 变量相关性

“VAR 变量和 WITH 变量之间的相关性”是指第一组中各变量和第二组中各变量的相关性，例如 X_1 和 Y_1 的相关系数为0.732 9，且多对变量的相关系数较高，说明适合利用典型相关进行分析。

2 典型相关系数检验

	典型相关	调整典型相关	近似标准误差	典型相关平方	特征值: Inv(E)*H = CanRsq/(1-CanRsq) 特征值	差分	比例	累积	H0 检验: 当前行和之后的所有行的典型相关都是零 似然比	近似 F 值	分子自由度	分母自由度	Pr > F
1	0.960413	0.951478	0.014169	0.922393	11.8854	9.6066	0.8122	0.8122	0.01604341	9.15	20	73.916	<.0001
2	0.833671	0.807183	0.055684	0.695007	2.2788	1.8207	0.1557	0.9679	0.20672553	4.15	12	61.144	<.0001
3	0.560496	0.516533	0.125218	0.314155	0.4581	0.4462	0.0313	0.9992	0.67780425	1.72	6	48	0.1374

输出 8-3 给出了典型相关系数及其检验。第一对典型变量的相关系数是0.960 4，调整后的典型相关系数是0.951 5，$p < 0.000\ 1$ 拒绝相关系数为零的原假设，这说明第一对典型相关系数在0.01 的显著性水平下显著。第二对典型变量的相关系数为0.833 7，调整后的典型相关系数为0.807 2，同样，第二对典型相关系数也通过检验。因此，我们选择前两组典型变量进行解释。

Multivariate Statistics and F Approximations

S=4 M=0 N=10

统计量	值	F 值	分子自由度	分母自由度	Pr > F
Wilks' Lambda	0.01604341	9.15	20	73.916	<.0001
Pillai's Trace	1.94327841	4.72	20	100	<.0001

输出 8-4 给出了典型相关系数是否为 0 的统计学检验，四种统计方法都拒绝了其为 0 的假设，因此可以认为典型相关系数是显著的。

3.根据下面两个表写出典型变量的标准化表达式。

VAR 变量 的标准化典型系数

	V1	V2	V3	V4
x1 人均地区生产总值	-0.1853	1.0392	-2.2279	1.5330
x2 人均可支配收入	-0.1330	1.6186	1.9103	-3.1056

WITH 变量 的标准化典型系数

	W1	W2	W3	W4
y1 每万从业人员有效发明专利数	0.3025	-0.0333	-1.7625	-0.7142
y2 每万从业人员发明专利申请数	0.1257	-0.3576	1.6355	-0.7237