第六章 聚类分析

一、农民生活水平的分类

1、16个地区农民生活水平的调查数据

地区 食品 衣着 燃料 住房 生活用品及其他 文化生活服务支出

北京 190.33 43.77 9.73 60.54 49.01 9.04

天津 135.20 36.40 10.47 44.16 36.49 3.94

河北 95.21 22.83 9.30 22.44 22.81 2.80

山西 104.78 25.11 6.40 9.89 18.17 3.25

内蒙古 128.41 27.63 8.94 12.58 23.99 3.27

辽宁 145.68 32.83 17.79 23.29 39.09 3.47

吉林 159.37 33.38 18.37 11.81 25.29 5.22

黑龙江 116.22 29.57 13.24 13.76 21.75 6.04

上海 221.11 38.64 12.53 115.65 50.82 5.89

江苏 144.98 29.12 11.67 42.60 27.30 5.74

浙江 169.92 32.75 12.72 47.12 34.35 5.00

安徽 153.11 23.09 15.62 23.54 18.18 6.39

福建 144.92 21.26 16.96 19.52 21.75 6.73

江西 140.54 21.50 17.64 19.19 15.97 4.49

山东 115.84 30.26 12.20 33.61 33.77 3.85

河南 101.18 23.26 8.46 20.20 20.50 4.30

2、cluster SAS程序

data chapter6;

input diqu x1 x2 x3 x4 x5 x6;

cards;

北京 190.33 43.77 9.73 60.54 49.01 9.04

天津 135.20 36.40 10.47 44.16 36.49 3.94

河北 95.21 22.83 9.30 22.44 22.81 2.80

山西 104.78 25.11 6.40 9.89 18.17 3.25

内蒙古 128.41 27.63 8.94 12.58 23.99 3.27

辽宁 145.68 32.83 17.79 23.29 39.09 3.47

吉林 159.37 33.38 18.37 11.81 25.29 5.22

黑龙江 116.22 29.57 13.24 13.76 21.75 6.04

上海 221.11 38.64 12.53 115.65 50.82 5.89

江苏 144.98 29.12 11.67 42.60 27.30 5.74

浙江 169.92 32.75 12.72 47.12 34.35 5.00

安徽 153.11 23.09 15.62 23.54 18.18 6.39

福建 144.92 21.26 16.96 19.52 21.75 6.73

江西 140.54 21.50 17.64 19.19 15.97 4.49

山东 115.84 30.26 12.20 33.61 33.77 3.85

河南 101.18 23.26 8.46 20.20 20.50 4.30

;

Proc cluster method=average;

Var x1 x2 x3 x4 x5 x6;

Run;

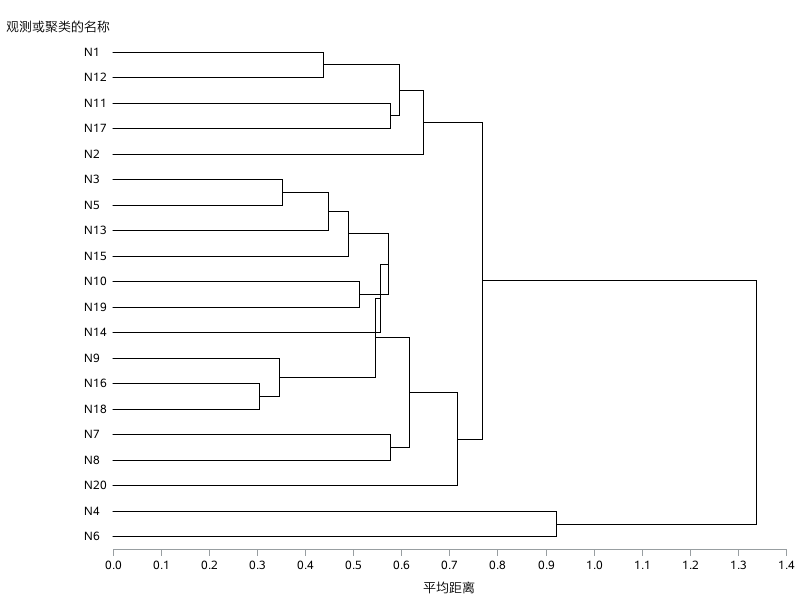
3、cluster SAS结果分析

| **协方差矩阵的特征值** | | | | |
| --- | --- | --- | --- | --- |
|  | **特征值** | **差分** | **比例** | **累积** |
| **1** | 1723.98909 | 1517.69878 | 0.8649 | 0.8649 |
| **2** | 206.29032 | 156.28979 | 0.1035 | 0.9684 |
| **3** | 50.00052 | 41.48797 | 0.0251 | 0.9935 |
| **4** | 8.51256 | 5.38257 | 0.0043 | 0.9977 |
| **5** | 3.12998 | 1.72124 | 0.0016 | 0.9993 |
| **6** | 1.40874 |  | 0.0007 | 1.0000 |

|  |  |
| --- | --- |
| **根均方总样本标准差** | 18.22695 |

|  |  |
| --- | --- |
| **观测之间的根均方距离** | 63.14002 |

| **聚类历史** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **聚类数** | **连接聚类** | | **频数** | **Norm RMS Distance** | **结值** |
| **15** | **OB3** | **OB16** | 2 | 0.111 |  |
| **14** | **OB13** | **OB14** | 2 | 0.1209 |  |
| **13** | **OB12** | **CL14** | 3 | 0.1918 |  |
| **12** | **OB5** | **OB8** | 2 | 0.2154 |  |
| **11** | **CL15** | **OB4** | 3 | 0.2289 |  |
| **10** | **OB2** | **OB10** | 2 | 0.2455 |  |
| **9** | **OB7** | **CL13** | 4 | 0.3436 |  |
| **8** | **OB6** | **CL9** | 5 | 0.3783 |  |
| **7** | **CL12** | **OB15** | 3 | 0.3973 |  |
| **6** | **CL11** | **CL7** | 6 | 0.4152 |  |
| **5** | **OB1** | **OB11** | 2 | 0.4903 |  |
| **4** | **CL10** | **CL8** | 7 | 0.503 |  |
| **3** | **CL4** | **CL6** | 13 | 0.6932 |  |
| **2** | **CL5** | **CL3** | 15 | 1.093 |  |
| **1** | **CL2** | **OB9** | 16 | 2.0475 |  |

 从图中可以看出，若分为两类，北京上海为一类，其他为一类，若分为四类，北京上海为一类，天津江苏山东浙江辽宁吉林为一类，黑龙江安徽福建江西为一类，河北河南山西内蒙古为一类；

若分为五类，则为(北京)、（上海）、（天津、山东、江苏、浙江、辽宁、吉林）、（黑龙江、安徽、福建、江西）、（河北、河南、山西、内蒙古）。

4、fastclus SAS程序

data chapter6;

input diqu x1 x2 x3 x4 x5 x6;

cards;

北京 190.33 43.77 9.73 60.54 49.01 9.04

天津 135.20 36.40 10.47 44.16 36.49 3.94

河北 95.21 22.83 9.30 22.44 22.81 2.80

山西 104.78 25.11 6.40 9.89 18.17 3.25

内蒙古 128.41 27.63 8.94 12.58 23.99 3.27

辽宁 145.68 32.83 17.79 23.29 39.09 3.47

吉林 159.37 33.38 18.37 11.81 25.29 5.22

黑龙江 116.22 29.57 13.24 13.76 21.75 6.04

上海 221.11 38.64 12.53 115.65 50.82 5.89

江苏 144.98 29.12 11.67 42.60 27.30 5.74

浙江 169.92 32.75 12.72 47.12 34.35 5.00

安徽 153.11 23.09 15.62 23.54 18.18 6.39

福建 144.92 21.26 16.96 19.52 21.75 6.73

江西 140.54 21.50 17.64 19.19 15.97 4.49

山东 115.84 30.26 12.20 33.61 33.77 3.85

河南 101.18 23.26 8.46 20.20 20.50 4.30

;

proc print;

run;

Proc fastclus maxclusters=4;

Run;

5、SAS输出结果

| **初始种子** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **聚类** | **diqu** | **x1** | **x2** | **x3** | **x4** | **x5** | **x6** |
| **1** | 0.0000000 | 141.6750000 | 29.4625000 | 12.6275000 | 32.4937500 | 28.7025000 | 4.9637500 |

二、女中学生体型聚类

1、8个体型变量的相关系数矩阵

| **Obs** | **x1** | **x3** | **x4** | **x5** | **x6** | **x7** | **x8** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | 0.846 | 0.805 | 0.859 | 0.478 | 0.398 | 0.301 | 0.382 |
| **2** | 1.000 | 0.881 | 0.826 | 0.376 | 0.326 | 0.277 | 0.415 |
| **3** | 0.881 | 1.000 | 0.801 | 0.380 | 0.319 | 0.237 | 0.345 |
| **4** | 0.826 | 0.801 | 1.000 | 0.436 | 0.329 | 0.327 | 0.365 |
| **5** | 0.376 | 0.380 | 0.436 | 1.000 | 0.762 | 0.730 | 0.629 |
| **6** | 0.362 | 0.319 | 0.329 | 0.762 | 1.000 | 0.583 | 0.557 |
| **7** | 0.277 | 0.237 | 0.327 | 0.730 | 0.583 | 1.000 | 0.539 |
| **8** | 0.415 | 0.345 | 0.365 | 0.629 | 0.577 | 0.539 | 1.000 |

2、SAS程序

data chapter6;

input x1 x2 x3 x4 x5 x6 x7 x8;

cards;

0.846 0.805 0.859 0.478 0.398 0.301 0.382

1.000 0.881 0.826 0.376 0.326 0.277 0.415

0.881 1.000 0.801 0.380 0.319 0.237 0.345

0.826 0.801 1.000 0.436 0.329 0.327 0.365

0.376 0.380 0.436 1.000 0.762 0.730 0.629

0.362 0.319 0.329 0.762 1.000 0.583 0.557

0.277 0.237 0.327 0.730 0.583 1.000 0.539

0.415 0.345 0.365 0.629 0.577 0.539 1.000

;

proc cluster method=average;

var x1 x2 x3 x4 x5 x6 x7 x8;

run;

3、SAS输出结果

| **协方差矩阵的特征值** | | | | |
| --- | --- | --- | --- | --- |
|  | **特征值** | **差分** | **比例** | **累积** |
| **1** | 0.60651497 | 0.56959482 | 0.9224 | 0.9224 |
| **2** | 0.03692015 | 0.02284652 | 0.0562 | 0.9786 |
| **3** | 0.01407363 | 0.01407363 | 0.0214 | 1.0000 |
| **4** | 0.00000000 | 0.00000000 | 0.0000 | 1.0000 |
| **5** | 0.00000000 | 0.00000000 | 0.0000 | 1.0000 |
| **6** | 0.00000000 | 0.00000000 | 0.0000 | 1.0000 |
| **7** | -.00000000 | 0.00000000 | -0.0000 | 1.0000 |
| **8** | -.00000000 |  | -0.0000 | 1.0000 |

|  |  |
| --- | --- |
| **根均方总样本标准差** | 0.286686 |

|  |  |
| --- | --- |
| **观测之间的根均方距离** | 1.146742 |

| **聚类历史** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **聚类数** | **连接聚类** | | **频数** | **Norm RMS Distance** | **结值** |
| **3** | **OB1** | **OB2** | 2 | 0.2676 |  |
| **2** | **OB3** | **OB4** | 2 | 0.419 |  |
| **1** | **CL3** | **CL2** | 4 | 1.1992 |  |