习题七

7.1 设$\left(\xi,\eta\right)\sim N\left(\mu_1,\delta_1^2;\mu_2,\delta_2^2;\rho\right)$，令 $\mu=\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}$, $\Sigma=\begin{bmatrix}\delta_1^2 & \rho\delta_1\delta_2\\ \rho\delta_1\delta_2 & \delta_2^2\end{bmatrix}$，于是二维正态分布$N\left(\mu_1,\delta_1^2;\mu_2,\delta_2^2;\rho\right)$可表示为$N_2\left(\mu,\Sigma\right)$.

（1）试证明$\left(\xi,\eta\right)$的联合密度函数$p\left(x_1,x_2\right)$可表示为

$$p\left(x_1,x_2\right)=\frac{1}{\left(2\pi\right)^{\frac{2}{2}}\left|\Sigma\right|^{\frac{1}{2}}}e^{-\frac{1}{2}\left(x-\mu\right)^T\Sigma^{-1}\left(x-\mu\right)},\text{其中 } \mathrm{x}=\left(x_1,x_2\right)^T$$

（2）设$\left(X_{(1)},X_{(2)},...,X_{(n)}\right)^T$为正态总体$\left(\xi,\ \eta\right)$的样本，证明样本均值

$\overline{X}\sim N_2\left(\mu,\frac{1}{n}\Sigma\right)$.

（1）证 因$\left(\xi,\eta\right)\sim N\left(\mu_1,\delta_1^2;\mu_2,\delta_2^2;\rho\right)$，故

$$P\left(x_1,x_2\right)=\frac{1}{2\pi\delta_1\delta_2\sqrt{1-\rho^2}}e^{-\frac{1}{2\left(1-\rho^2\right)}\left[\left(\frac{x_1-x_2}{\delta_1}\right)^2-2\rho\frac{\left(x_1-\mu_1\right)\left(x_2-\mu_2\right)}{\delta_1\delta_2}+\left(\frac{x_2-\mu_2}{\delta_2}\right)^2\right]}$$

$$=\frac{1}{\left(2\pi\right)^{\frac{3}{2}}\sqrt{\delta_1^2\delta_2^2\left(1-\rho^2\right)}}e^{-\frac{1}{2}\begin{pmatrix}x_1-\mu_1\\ x_2-\mu_2\end{pmatrix}^T\begin{bmatrix}\delta_1^2 & \rho\delta_1\delta_2\\ \rho\delta_1\delta_2 & \delta_2^2\end{bmatrix}^{-1}\begin{pmatrix}x_1-\mu_1\\ x_2-\mu_2\end{pmatrix}}$$

$$=\frac{1}{\left(2\pi\right)^{\frac{3}{2}}\left|\Sigma\right|^{\frac{1}{2}}}e^{-\frac{1}{2}\left(x-\mu\right)^T\Sigma^{-1}\left(x-\mu\right)}$$

（2）二维正态总体，样本均值仍服从正态分布，而

$$E\overline{X}=\begin{pmatrix}E\overline{X_1}\\ E\overline{X_2}\end{pmatrix}=\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}=\mu$$

$$D\overline{X}=E\left(\overline{X}-E\overline{X}\right)\left(\overline{X}-E\overline{X}\right)^T=E\left(\overline{X}-\mu\right)\left(\overline{X}-\mu\right)^T$$

$$=E\begin{pmatrix}\overline{X_1}-\mu_1\\ \overline{X_2}-\mu_2\end{pmatrix}\begin{pmatrix}\overline{X_1}-\mu_1 & \overline{X_2}-\mu_2\end{pmatrix}=E\begin{pmatrix}\left(\overline{X_1}-\mu_1\right)^2 & \left(\overline{X_1}-\mu_1\right)\left(\overline{X_2}-\mu_2\right)\\ \left(\overline{X_1}-\mu_1\right)\left(\overline{X_2}-\mu_2\right) & \left(\overline{X_2}-\mu_2\right)^2\end{pmatrix}$$

$$= \begin{pmatrix} D\overline{X_1} & \operatorname{cov}\left(\overline{X_1}, \overline{X_2}\right) \\ \operatorname{cov}\left(\overline{X_1}, \overline{X_2}\right) & D\overline{X_2} \end{pmatrix} = \frac{1}{n}\begin{pmatrix} \delta_1^2 & \rho\delta_1\delta_2 \\ \rho\delta_1\delta_2 & \delta_2^2 \end{pmatrix} = \frac{1}{n}\Sigma$$

故 $\overline{X} \sim N\left(\mu, \frac{1}{n}\Sigma\right)$.

7.2 随机抽取某班级四名同学的数学、物理和化学三门课程的期中考试成绩，结果如下：

	数学	物理	化学
甲	70	75	65
已	60	70	50
丙	80	75	70
丁	90	80	80

（1）写出样本数据阵 X；

（2）求出样本均值 $\overline{X}$，样本协方差阵 S，样本相关阵 R；

（3）分别把数学成绩 x 极差标准化为 $\tilde{x}$，把物理成绩 y 标准差标准化 $\tilde{y}$；

（4）写出甲乙两同学的考试成绩的马氏距离表达式，并求出甲乙两同学考试成绩的欧氏距离.

解 (1) 样本数据阵为: $X = \begin{bmatrix} 70 & 75 & 65 \\ 60 & 70 & 50 \\ 80 & 75 & 70 \\ 90 & 80 & 80 \end{bmatrix}$;

(2) 样本协方差阵(对称矩阵)为: $S = \begin{bmatrix} 125 & & \\ 37.5 & 12.5 & \\ 118.75 & 37.5 & 117.1875 \end{bmatrix}$,

样本相关阵(对称矩阵)为: $R = \begin{bmatrix} 1 & & \\ 0.948683298 & 1 & \\ 0.981155781 & 0.979795897 & 1 \end{bmatrix}$;

(3) 数学成绩的极差标准化 $\tilde{x} = \begin{bmatrix} (70-75)/30 \\ (60-75)/30 \\ (80-75)/30 \\ (90-75)/30 \end{bmatrix} = \begin{bmatrix} -1/6 \\ -1/2 \\ 1/6 \\ 1/2 \end{bmatrix} = \begin{bmatrix} -1.6667 \\ -0.5 \\ 1.6667 \\ 0.5 \end{bmatrix}$,

物理成绩 y 标准差标准化 $\tilde{y} = \begin{bmatrix} (75-75)/4.0825 \\ (70-75)/4.0825 \\ (75-75)/4.0825 \\ (80-75)/4.0825 \end{bmatrix} = \begin{bmatrix} 0 \\ -1.2247 \\ 0 \\ 1.2247 \end{bmatrix}$;

(4) 甲乙两同学的考试成绩的马氏距离表达式:

$$d(X_{(1)}^T,X_{(2)}^T)=\sqrt{(X_{(1)}-X_{(2)})^T S^{-1}(X_{(1)}-X_{(2)})}$$

$$=\sqrt{[10,5,15]\begin{bmatrix}125 & & \\ 37.5 & 12.5 & \\ 118.75 & 37.5 & 117.1875\end{bmatrix}^{-1}\begin{bmatrix}10\\5\\15\end{bmatrix}}$$

甲乙两同学考试成绩的欧氏距离为:

$$d(X_{(1)}^T,X_{(2)}^T)=\sqrt{(X_{(1)}-X_{(2)})^T(X_{(1)}-X_{(2)})}$$

$$=\sqrt{[10,5,15]\begin{bmatrix}10\\5\\15\end{bmatrix}}=\sqrt{100+25+225}\approx 18.7083$$

7.3 设对 m 维随机变量进行 n 次观测得到的样本数据阵为 X, $X=\left(x_{ij}\right)_{n\times m}$，令 $\tilde{X}$ 为标准差标准化之后的样本数据阵即

$$\tilde{X}=\left(\tilde{x}_{ij}\right)_{n\times m}\text{，其中}\tilde{x}_{ij}=\frac{x_{ij}-\overline{x}_j}{s_j}\left(i=1,2,\ldots,n;\ j=1,2,\ldots,m\right),$$

试证明样本相关阵 $R=\dfrac{1}{n-1}\tilde{X}^T\tilde{X}$.

证 $\dfrac{1}{n-1}\tilde{X}^T\tilde{X}$ 的第 i 行第 j 列元素为 $\tilde{X}^T$ 的第 i 行与 $\tilde{X}$ 的第 j 列对应元素相乘再相加的 $\dfrac{1}{n-1}$ 倍。即 $\tilde{X}$ 的第 i 列与第 j 列对应元素乘积之和的 $\dfrac{1}{n-1}$ 倍。用公式表示即为

$$\frac{1}{n-1}\sum_{k=1}^{n}\tilde{x}_{ki}\tilde{x}_{kj}=\frac{1}{n-1}\sum_{k=1}^{n}\frac{x_{ki}-\overline{x}_i}{s_i}\frac{x_{kj}-\overline{x}_j}{s_j}$$

$$=\frac{1}{n-1}\sum_{k=1}^{n}\frac{x_{ki}-\overline{x}_i}{\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(x_{ki}-\overline{x}_i\right)^2}}\frac{x_{kj}-\overline{x}_j}{\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(x_{kj}-\overline{x}_j\right)^2}}$$

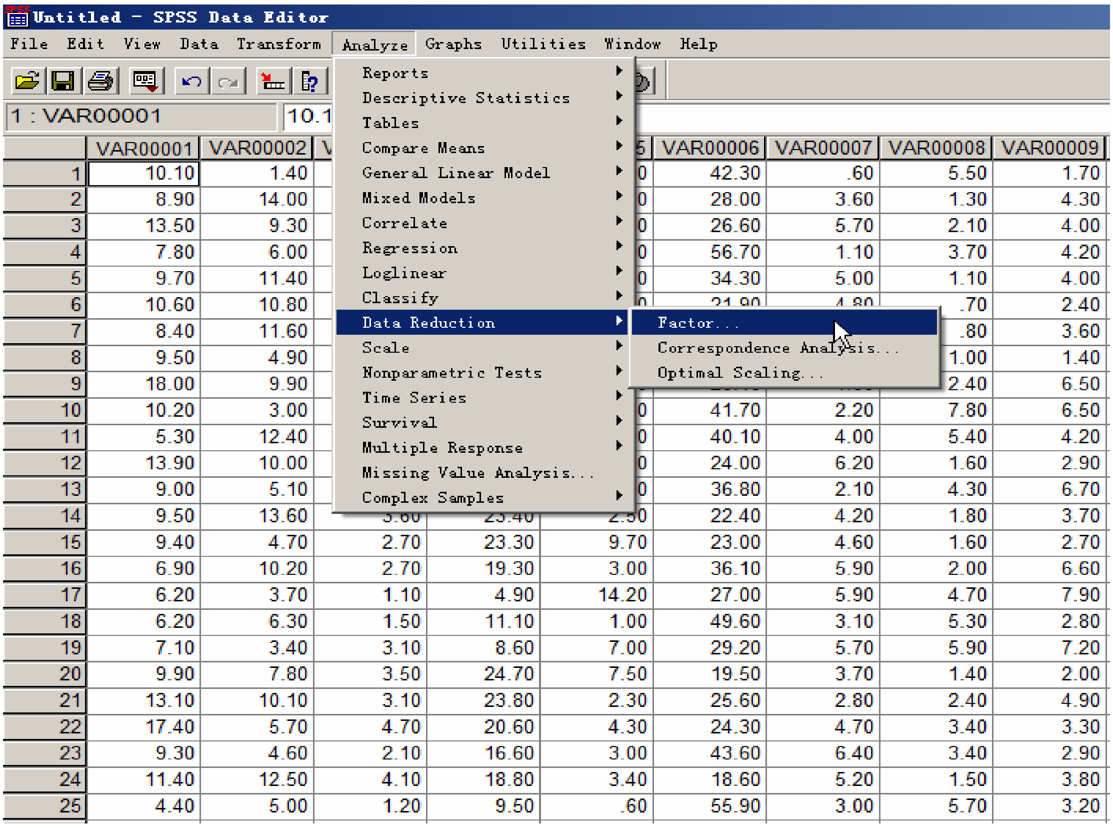
$$=\frac{\sum_{k=1}^{n}\left(x_{ki}-\overline{x}_i\right)\left(x_{kj}-\overline{x}_j\right)}{\sqrt{\sum_{k=1}^{n}\left(x_{ki}-\overline{x}_i\right)^2}\sqrt{\sum_{k=1}^{n}\left(x_{kj}-\overline{x}_j\right)^2}}$$

上式正是样本数据阵 X 的第 i 列与第 j 列的样本相关系数，故

$$R=\frac{1}{n-1}\tilde{X}^T\tilde{X}.$$

7.4 试借助统计分析工具（SPSS, SAS, R），对 7.2 节的例 3 进行主成分分析.

解 利用统计分析软件 SPSS 的操作结果截图如下



Communalities

	Initial	Extraction
VAR00001	1.000	.472
VAR00002	1.000	.917
VAR00003	1.000	.769
VAR00004	1.000	.795
VAR00005	1.000	.874
VAR00006	1.000	.867
VAR00007	1.000	.624
VAR00008	1.000	.745
VAR00009	1.000	.706

Extraction Method: Principal Component Analysis.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.006	44.516	44.516	4.006	44.516	44.516
2	1.635	18.167	62.683	1.635	18.167	62.683
3	1.128	12.532	75.215	1.128	12.532	75.215
4	.955	10.607	85.822			

Extraction Method: Principal Component Analysis.

Component Matrix(a)

	Component		
	1	2	3
VAR00001	.606	-.072	-.316
VAR00002	.622	-.303	.663
VAR00003	.854	-.045	.193
VAR00004	.756	-.236	-.410
VAR00005	.272	.827	-.341
VAR00006	-.876	-.299	.102
VAR00007	.595	.451	.258
VAR00008	-.841	.183	-.058
VAR00009	-.221	.686	.433

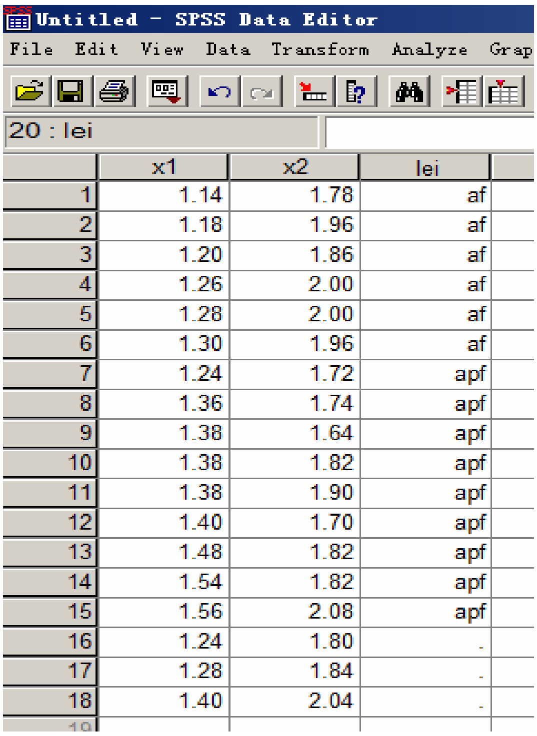
Extraction Method: Principal Component Analysis.

a 3 components extracted.

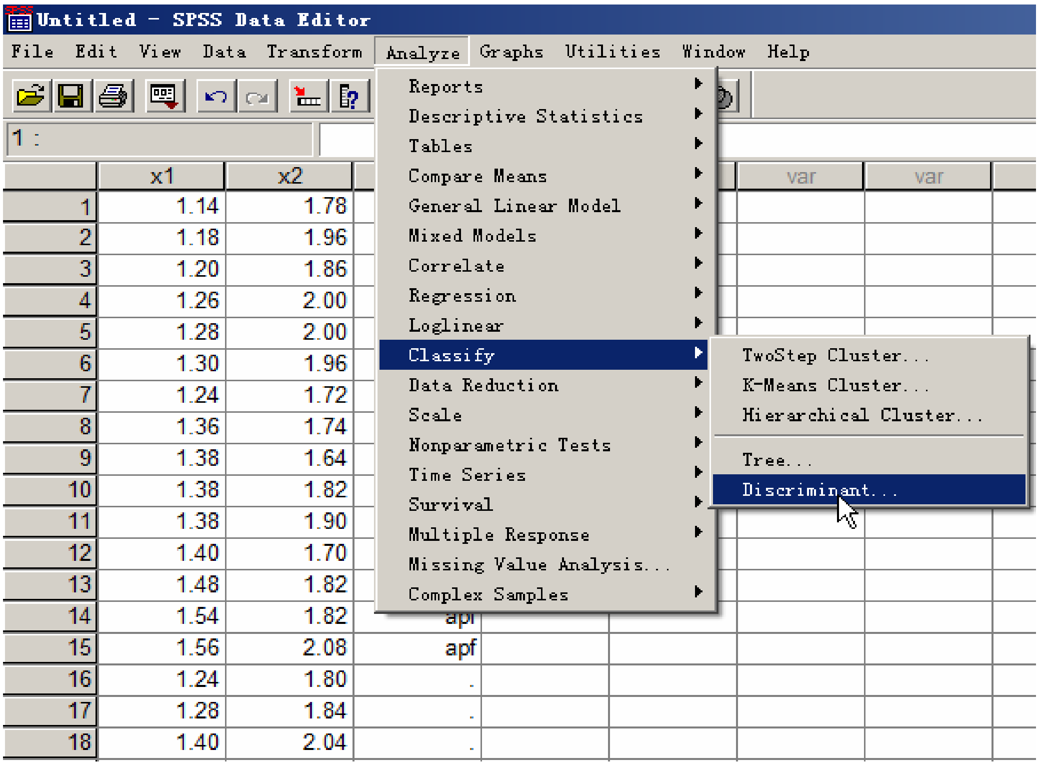
7.5 试借助统计分析工具对 7.3 节中关于蠓的分类一例中的数据进行判别分析.

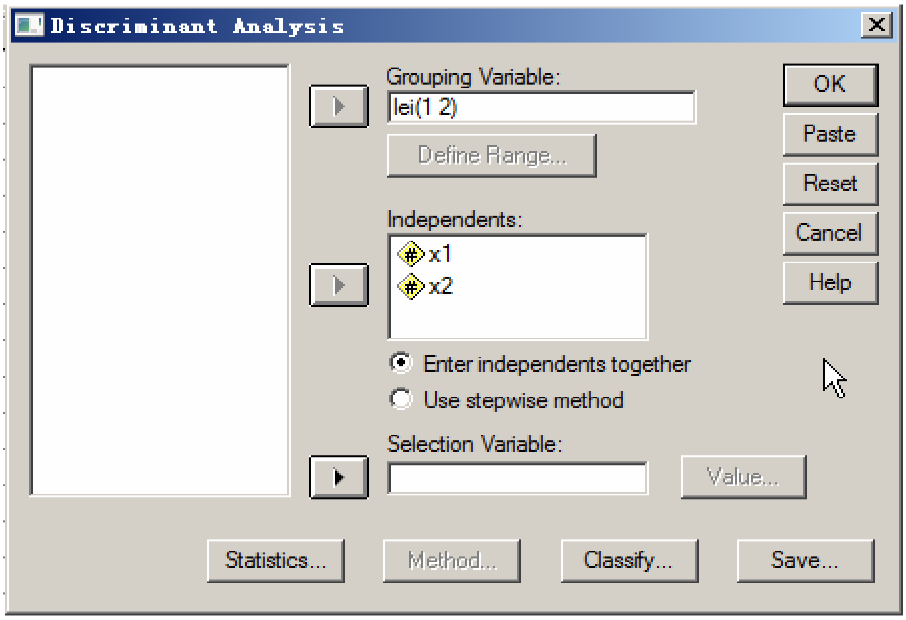
解 利用统计分析软件 SPSS 的操作结果截图如下

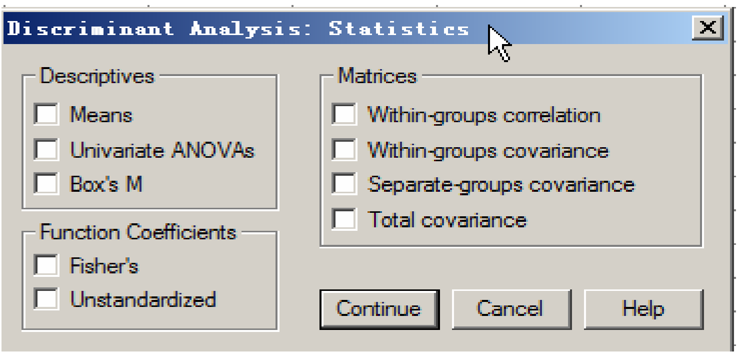
先建立 SPSS 数据文件

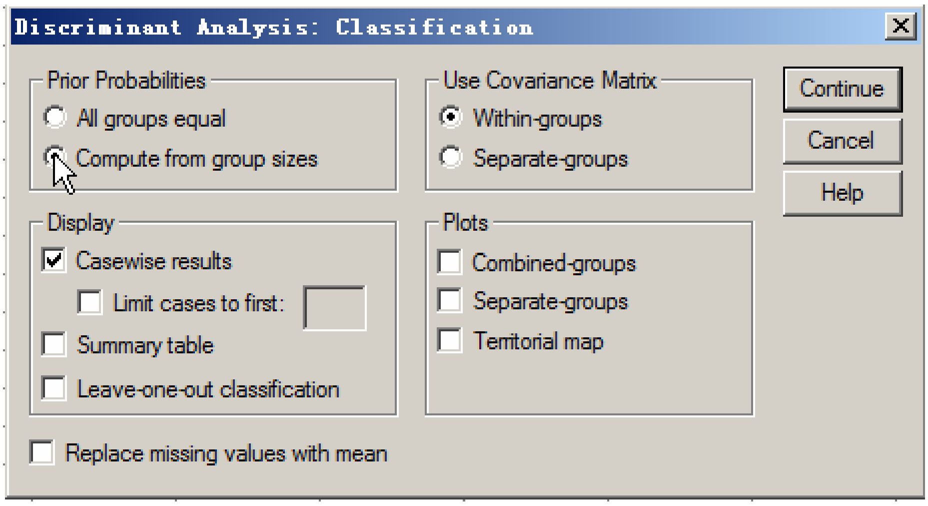


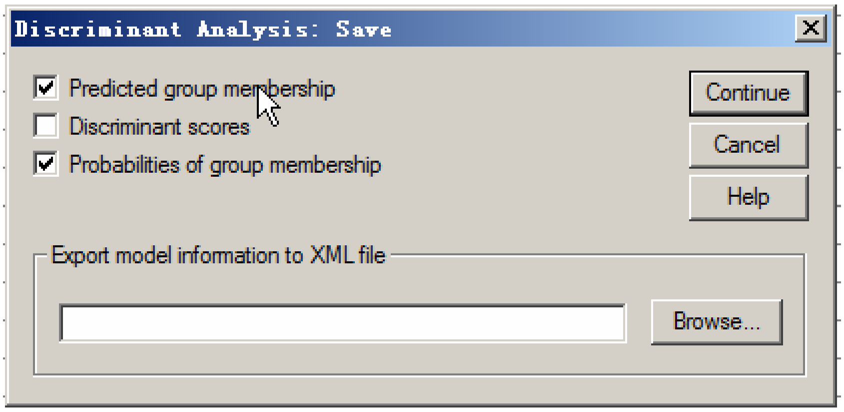
	x1	x2	lei
1	1.14	1.78	af
2	1.18	1.96	af
3	1.20	1.86	af
4	1.26	2.00	af
5	1.28	2.00	af
6	1.30	1.96	af
7	1.24	1.72	apf
8	1.36	1.74	apf
9	1.38	1.64	apf
10	1.38	1.82	apf
11	1.38	1.90	apf
12	1.40	1.70	apf
13	1.48	1.82	apf
14	1.54	1.82	apf
15	1.56	2.08	apf
16	1.24	1.80	.
17	1.28	1.84	.
18	1.40	2.04	.

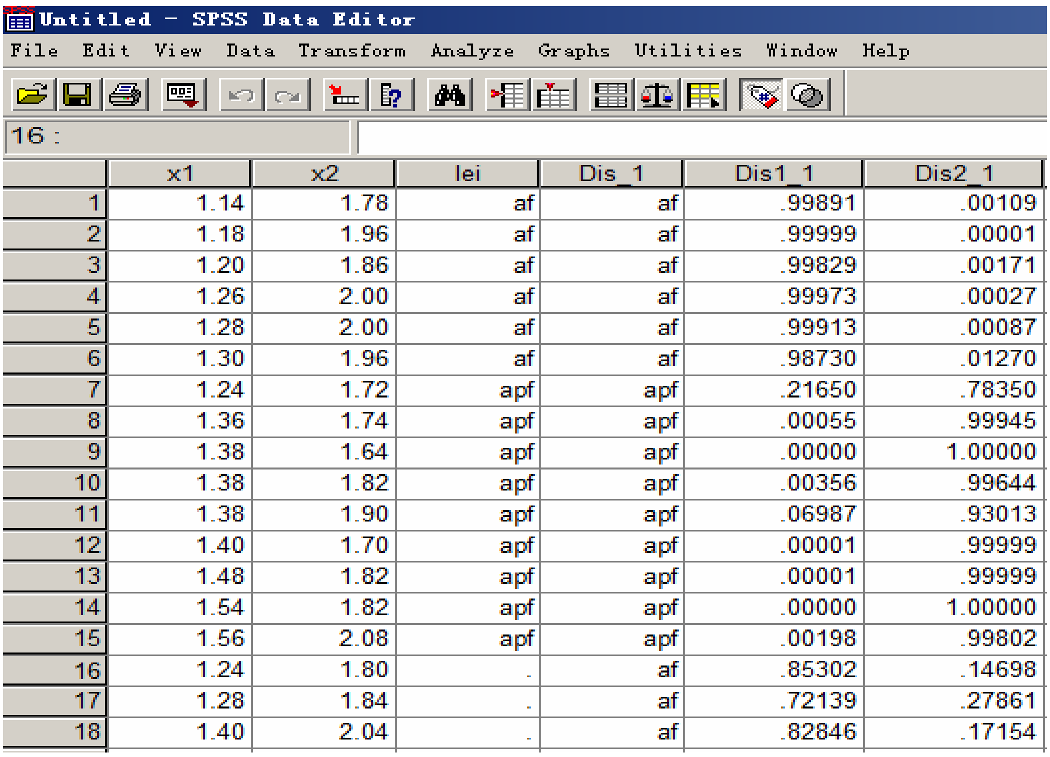

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Complex Samples
TwoStep Cluster...
K-Means Cluster...
Hierarchical Cluster...
Tree...
Discriminant...
x1 x2 var var
1 1.14 1.78
2 1.18 1.96
3 1.20 1.86
4 1.26 2.00
5 1.28 2.00
6 1.30 1.96
7 1.24 1.72
8 1.36 1.74
9 1.38 1.64
10 1.38 1.82
11 1.38 1.90
12 1.40 1.70
13 1.48 1.82
14 1.54 1.82
15 1.56 2.08 apf
16 1.24 1.80
17 1.28 1.84
18 1.40 2.04


Discriminant Analysis
Grouping Variable:
lei(1 2)
Define Range...
Independents:
x1
x2
Enter independents together
Use stepwise method
Selection Variable:
Value...
OK
Paste
Reset
Cancel
Help
Statistics...
Method...
Classify...
Save...


Discriminant Analysis: Statistics
Descriptives
Means
Univariate ANOVAs
Box's M
Function Coefficients
Fisher's
Unstandardized
Matrices
Within-groups correlation
Within-groups covariance
Separate-groups covariance
Total covariance
Continue
Cancel
Help





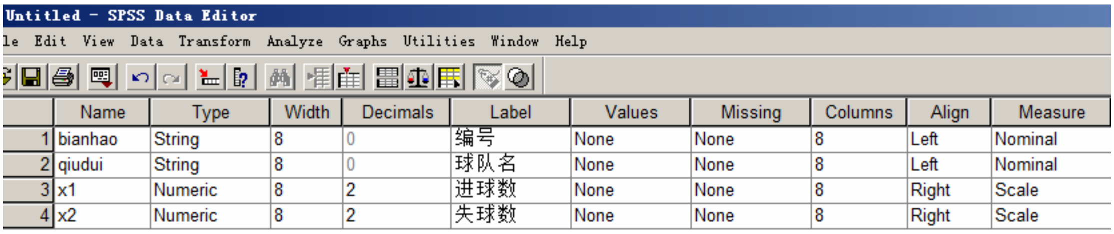


	x1	x2	lei	Dis_1	Dis1_1	Dis2_1
1	1.14	1.78	af	af	.99891	.00109
2	1.18	1.96	af	af	.99999	.00001
3	1.20	1.86	af	af	.99829	.00171
4	1.26	2.00	af	af	.99973	.00027
5	1.28	2.00	af	af	.99913	.00087
6	1.30	1.96	af	af	.98730	.01270
7	1.24	1.72	apf	apf	.21650	.78350
8	1.36	1.74	apf	apf	.00055	.99945
9	1.38	1.64	apf	apf	.00000	1.00000
10	1.38	1.82	apf	apf	.00356	.99644
11	1.38	1.90	apf	apf	.06987	.93013
12	1.40	1.70	apf	apf	.00001	.99999
13	1.48	1.82	apf	apf	.00001	.99999
14	1.54	1.82	apf	apf	.00000	1.00000
15	1.56	2.08	apf	apf	.00198	.99802
16	1.24	1.80	.	af	.85302	.14698
17	1.28	1.84	.	af	.72139	.27861
18	1.40	2.04	.	af	.82846	.17154

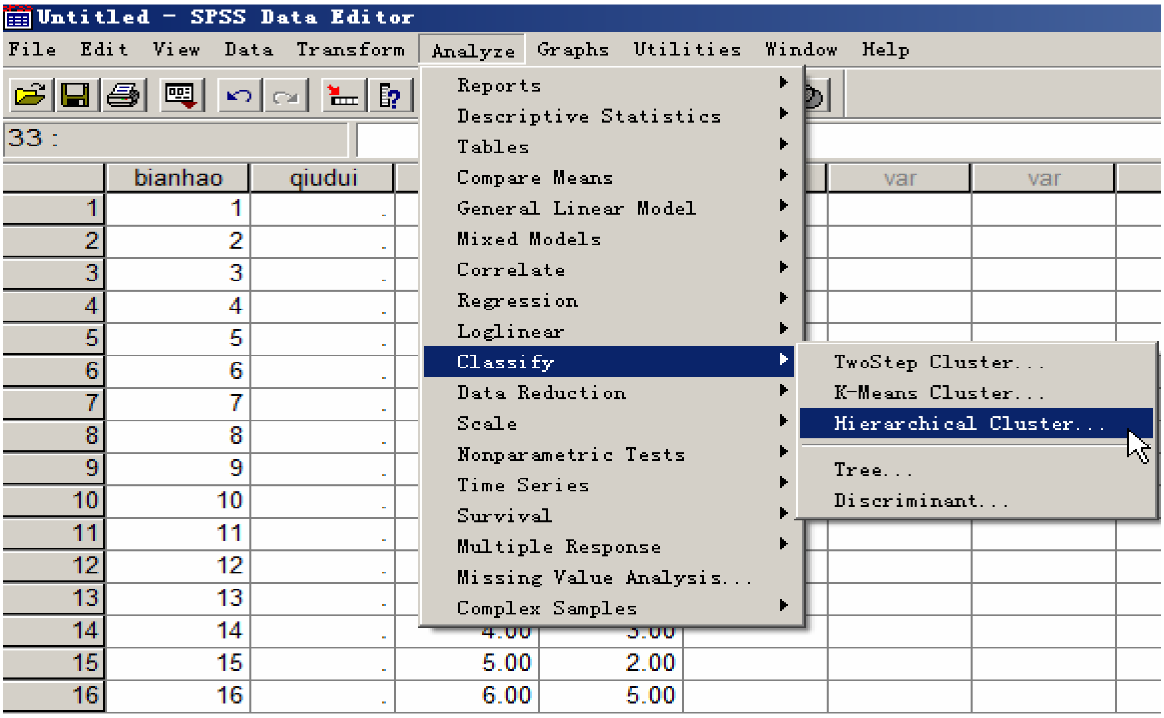
根据上述分析结果知,若采用总的样本协方差阵的方法的判别结果为: 第 16,17,18 号样品是第一类(为 af),他们为第一类的后验概率分别为 0.85302,0.72139,0.82846

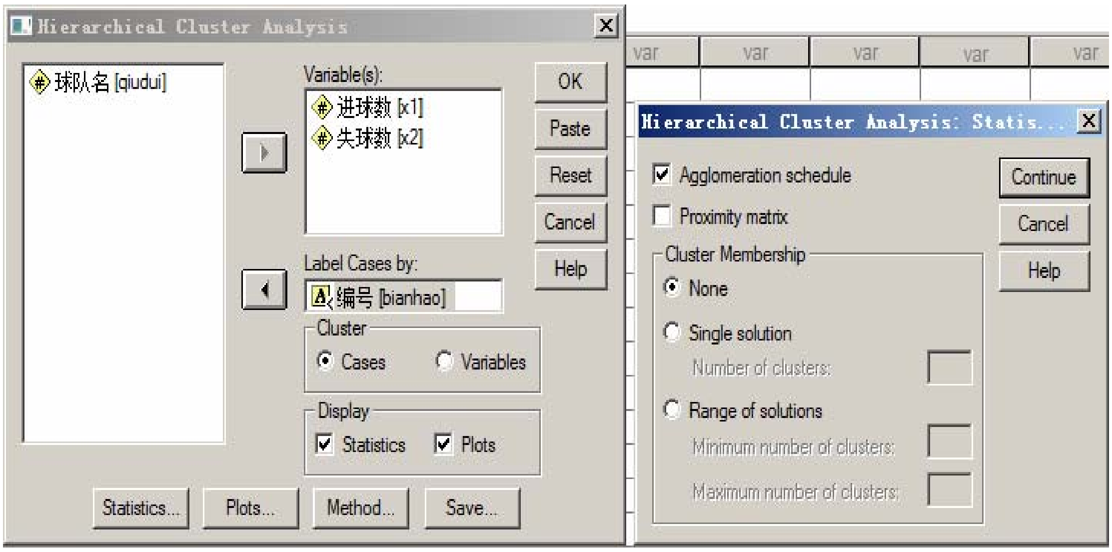
7.6 试借助统计分析工具验算 7.4 节中 2002 年足球世界杯 16 强的系统聚类的结果.

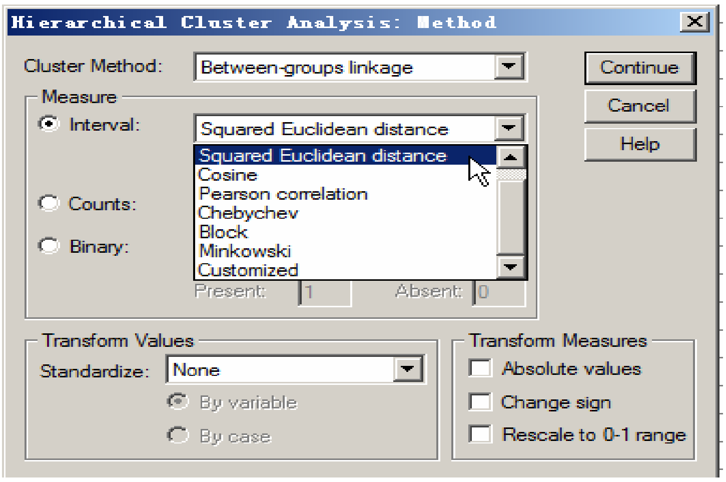
解 首先建立 SPSS 数据文件

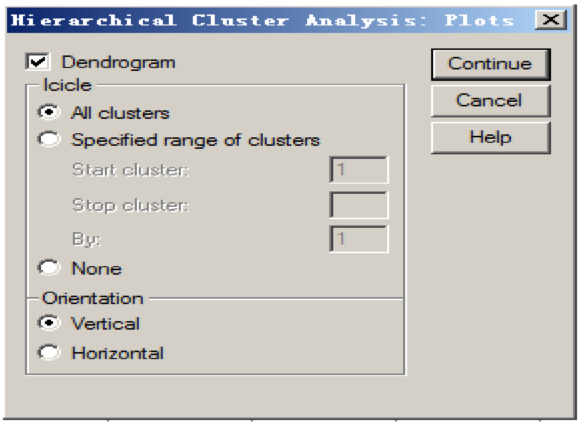


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	bianhao	String	8	0	编号	None	None	8	Left	Nominal
2	qiudui	String	8	0	球队名	None	None	8	Left	Nominal
3	x1	Numeric	8	2	进球数	None	None	8	Right	Scale
4	x2	Numeric	8	2	失球数	None	None	8	Right	Scale









系统聚类的结果用谱系图表示如下

```
* * * * * * H I E R A R C H I C A L   C L U S T E R   A N A L Y S I S * * * * * *

 Dendrogram using Average Linkage (Between Groups)

                         Rescaled Distance Cluster Combine

    C A S E        0         5        10        15        20        25
  Label      Num  +---------+---------+---------+---------+---------+

      10      10   ┐
      15      15   ┤
       1       1   ┼┐
      13      13   ┘│
      11      11   ┐│
      14      14   ┼┼────┐
       6       6   ┘│    ├──┐
       7       7   ─┘    │   │
```

```
12    12
 4     4
16    16
 8     8
 2     2
 5     5
 9     9
 3     3
```

7.7 据调查市场上销售的 9 种饮料的热量，咖啡因含量，钠含量及价格的数据如下：

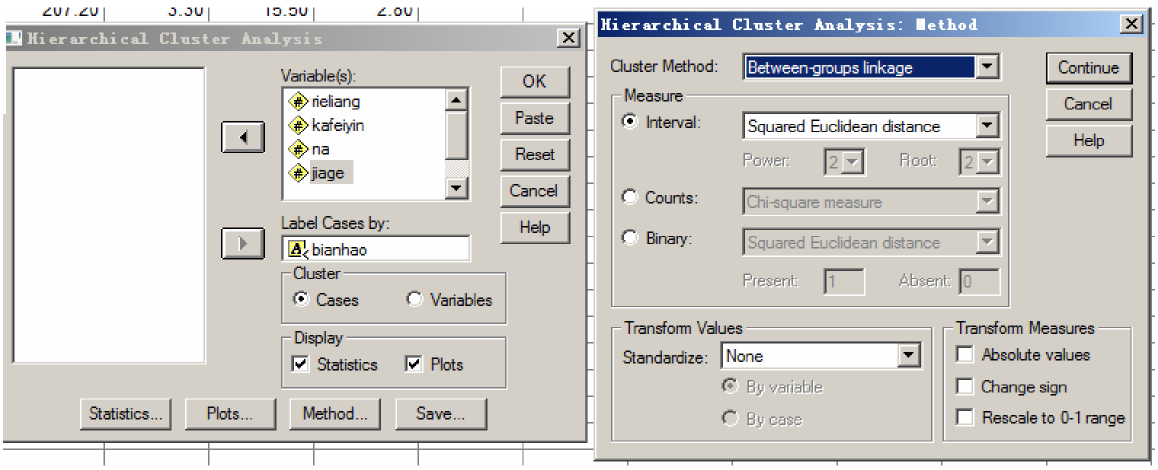
饮料编号	热量	咖啡因含量	钠含量	价格
1	207.2	3.3	15.5	2.8
2	36.8	5.9	12.9	3.3
3	72.2	7.3	8.2	2.4
4	36.7	0.4	10.5	4
5	121.7	4.1	9.2	3.5
6	89.1	4	10.2	3.3
7	146.7	4.3	9.7	1.8
8	57.6	2.2	13.6	2.1
9	95.9	0	8.5	1.3

（1）试借助统计分析软件对 9 种饮料进行系统聚类.

（2）借助统计分析工具对数进行主成分分析.

（3）根据（1）中分为三个类的聚类结果，判别一个未知类别的样品 $(38.5, 3.7, 7.7, 7.2)^T$ 属于其中哪一个类.

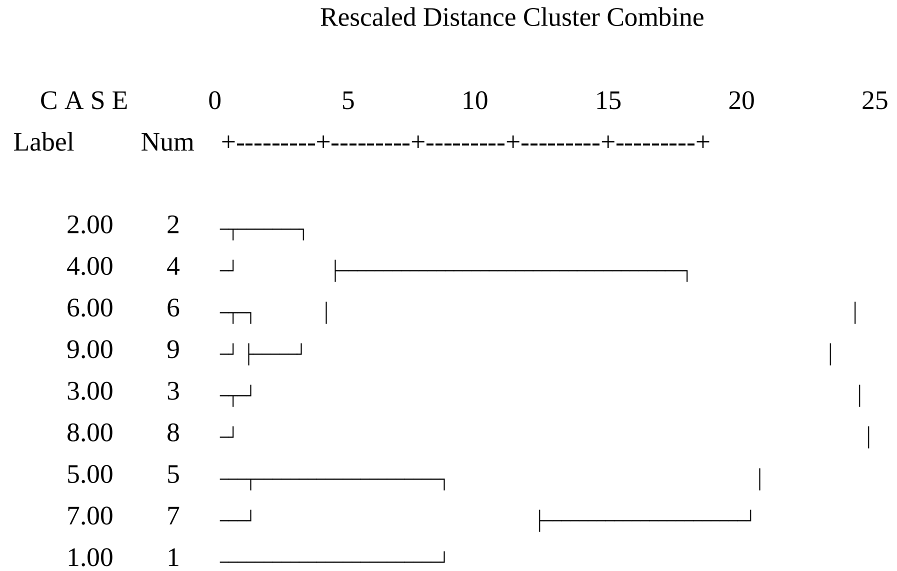
解 （1）系统聚类



聚类结果的谱系图：

* * * * * * H I E R A R C H I C A L   C L U S T E R   A N A L Y S I S * * * * * *

Dendrogram using Average Linkage (Between Groups)



当指定聚成三个类时：

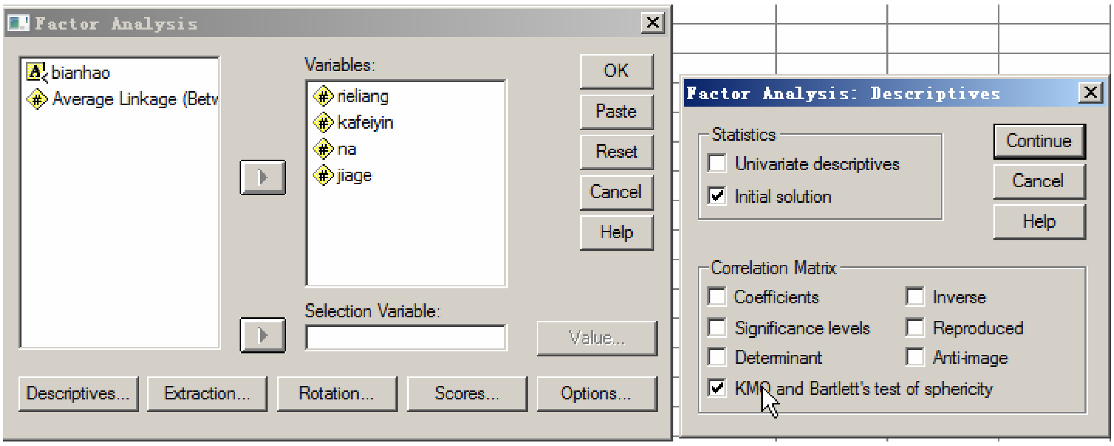
Untitled - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

9 :

	bianhao	rieliang	kafeiyin	na	jiage	CLU3_1
1	1.00	207.20	3.30	15.50	2.80	1
2	2.00	36.80	5.90	12.90	3.30	2
3	3.00	72.20	7.30	8.20	2.40	2
4	4.00	36.70	.40	10.50	4.00	2
5	5.00	121.70	4.10	9.20	3.50	3
6	6.00	89.10	4.00	10.20	3.30	2
7	7.00	146.70	4.30	9.70	1.80	3
8	8.00	57.60	2.20	13.60	2.10	2
9	9.00	95.90	.00	8.50	1.30	2

（2）作主成分分析时要求变元间具有较强的相关性，所以，第一步对原始数据是否适合作主成分分析要进行检验：

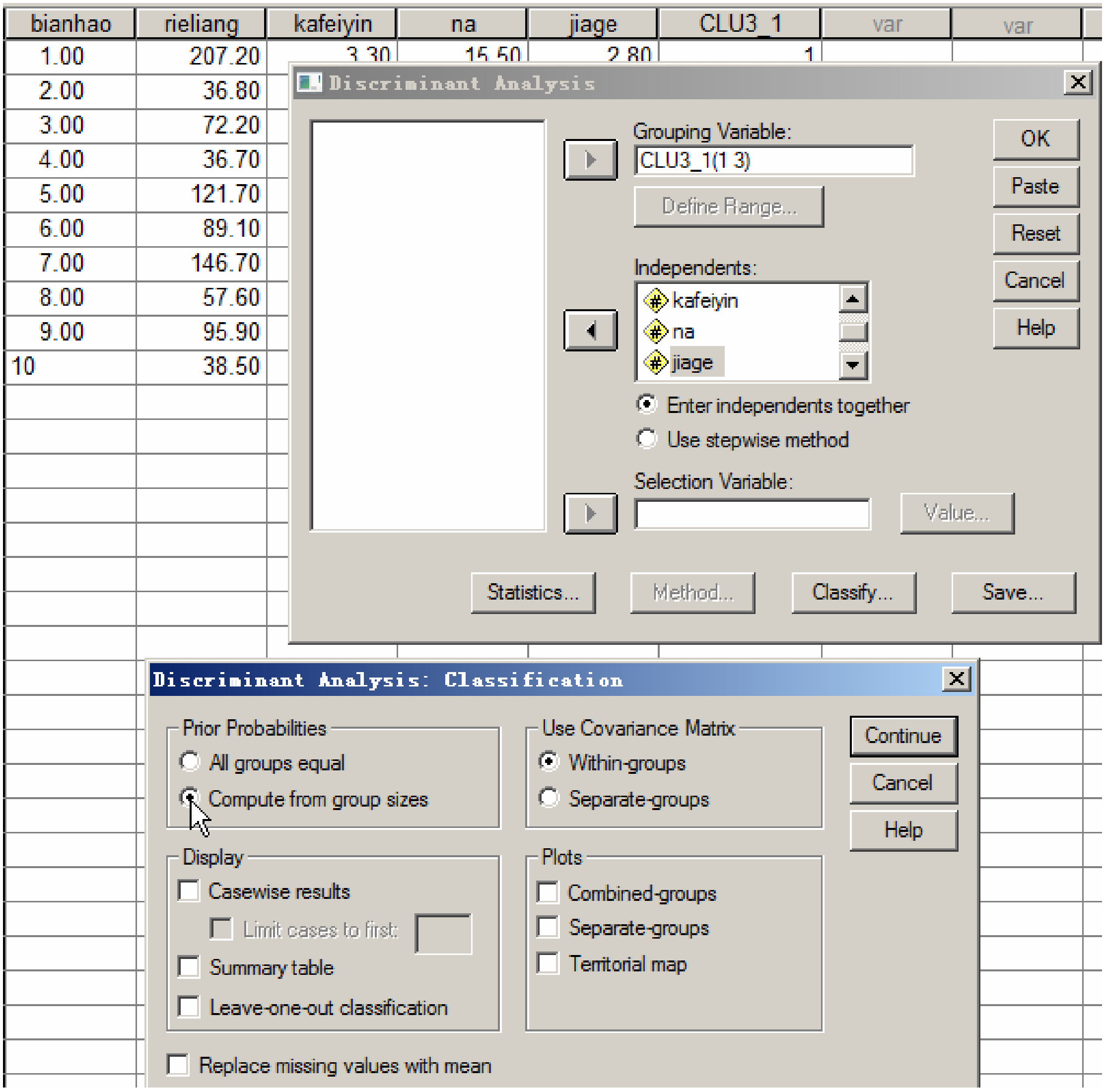


KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.360
Bartlett's Test of Sphericity	Approx. Chi-Square	1.296
	df	6
	Sig.	.972

检验的P值SIG=0.972，说明原变量具有弱相关性,不适合提取主成分

(3) 判别分析结果如下：



bianhao	rieliang	kafeiyin	na	jiage	CLU3_1	Dis_1	Dis1_1	Dis2_1	Dis3_1
1.00	207.20	3.30	15.50	2.80	1	1	1.00000	.00000	.00000
2.00	36.80	5.90	12.90	3.30	2	2	.00000	1.00000	.00000
3.00	72.20	7.30	8.20	2.40	2	2	.00000	1.00000	.00000
4.00	36.70	.40	10.50	4.00	2	2	.00000	1.00000	.00000
5.00	121.70	4.10	9.20	3.50	3	3	.00000	.00027	.99973
6.00	89.10	4.00	10.20	3.30	2	2	.00000	.80500	.19500
7.00	146.70	4.30	9.70	1.80	3	3	.00000	.00009	.99991
8.00	57.60	2.20	13.60	2.10	2	2	.00000	.99999	.00001
9.00	95.90	.00	8.50	1.30	2	2	.00000	.99999	.00001
10	38.50	3.70	7.70	2.00	.	2	.00000	1.00000	.00000

即这个(编号为 10 的)样品属第二类(后验概率约为 1)