# 数学建模 B 作业：多元统计作业（R 版）

## 1 回归分析

| x1 | x2 | x3 | x4 | y |
|---|---|---|---|---|
| 7 | 26 | 6 | 20 | 78.5 |
| 1 | 29 | 15 | 6 | 74.3 |
| 11 | 56 | 8 | 32 | 104.3 |
| 11 | 31 | 8 | 47 | 87.6 |
| 7 | 52 | 6 | 33 | 95.9 |
| 11 | 55 | 9 | 22 | 109.2 |
| 3 | 71 | 17 | 26 | 102.7 |
| 1 | 31 | 22 | 44 | 72.5 |
| 2 | 54 | 18 | 22 | 93.1 |
| 21 | 47 | 4 | 50 | 115.9 |
| 1 | 40 | 23 | 9 | 83.8 |
| 11 | 66 | 9 | 12 | 113.3 |
| 10 | 68 | 8 | 12 | 109.4 |

（1） 先用逐步回归的方法进行变量筛选，得到模型检验 x3,x4 系数不显著，

```
> summary(lm.both)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3051 -0.7536 -0.3132  0.4005  3.3467

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 49.51753    3.52062  14.065 6.34e-07 ***
x1           1.90001    0.21020   9.039 1.79e-05 ***
x2           0.62660    0.04217  14.861 4.14e-07 ***
x3           0.33904    0.16952   2.000   0.0805 .
x4          -0.09450    0.04996  -1.891   0.0952 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.039 on 8 degrees of freedom
Multiple R-squared:  0.9878,    Adjusted R-squared:  0.9816
F-statistic: 161.4 on 4 and 8 DF,  p-value: 1.112e-07
```

```
Start:  AIC=71.44
y ~ 1

       Df Sum of Sq     RSS    AIC
+ x2    1   1809.43  906.34 59.178
+ x1    1   1450.08 1265.69 63.519
+ x3    1    776.36 1939.40 69.067
<none>               2715.76 71.444
+ x4    1     15.69 2700.07 73.369

Step:  AIC=59.18
y ~ x2

       Df Sum of Sq     RSS    AIC
+ x1    1    848.43   57.90 25.420
+ x3    1    490.89  415.44 51.037
+ x4    1    161.74  744.59 58.623
<none>               906.34 59.178
- x2    1   1809.43 2715.76 71.444

Step:  AIC=25.42
y ~ x2 + x1

       Df Sum of Sq     RSS    AIC
+ x3    1      9.79   48.11 25.011
<none>               57.90 25.420
+ x4    1      8.04   49.87 25.477
- x1    1    848.43  906.34 59.178
- x2    1   1207.78 1265.69 63.519
```

```
Step:  AIC=25.01
y ~ x2 + x1 + x3

       Df Sum of Sq     RSS    AIC
+ x4    1     14.87   33.24 22.206
<none>               48.11 25.011
- x3    1      9.79   57.90 25.420
- x1    1    367.33  415.44 51.037
- x2    1   1178.96 1227.07 65.117

Step:  AIC=22.21
y ~ x2 + x1 + x3 + x4

       Df Sum of Sq    RSS    AIC
<none>              33.24 22.206
- x4    1     14.87  48.11 25.011
- x3    1     16.62  49.87 25.477
- x1    1    339.53 372.78 51.628
- x2    1    917.73 950.98 63.803
```

（2） 根据逐步回归顺序，建立 y 与 x1,x2,x3 的回归模型，x3 系数不显著，

```
> summary(lm1)

Call:
lm(formula = y ~ x1 + x2 + x3, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2543 -1.4726  0.1755  1.5409  3.9711

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 48.19363    3.91330  12.315 6.17e-07 ***
x1           1.69589    0.20458   8.290 1.66e-05 ***
x2           0.65691    0.04423  14.851 1.23e-07 ***
x3           0.25002    0.18471   1.354    0.209
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.312 on 9 degrees of freedom
Multiple R-squared:  0.9823,    Adjusted R-squared:  0.9764
F-statistic: 166.3 on 3 and 9 DF,  p-value: 3.367e-08
```

（3） 建立 y 与 x1,x2 的回归模型，系数检验显著。建立线性回归方程为：

$y = 52.58 + 1.47x_1 + 0.66x_2$，且拟合优度达到 R2=0.9787。可知，方程拟合效果很好。F Value=229.50，Pr > F 远小于 0.05，故回归方程的线性及各参数的显著性检验均通过。

```
> summary(lm2)

Call:
lm(formula = y ~ x1 + x2, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
-2.893 -1.574 -1.302  1.363  4.048

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.57735    2.28617   23.00 5.46e-10 ***
x1           1.46831    0.12130   12.11 2.69e-07 ***
x2           0.66225    0.04585   14.44 5.03e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.406 on 10 degrees of freedom
Multiple R-squared:  0.9787,    Adjusted R-squared:  0.9744
F-statistic: 229.5 on 2 and 10 DF,  p-value: 4.407e-09
```

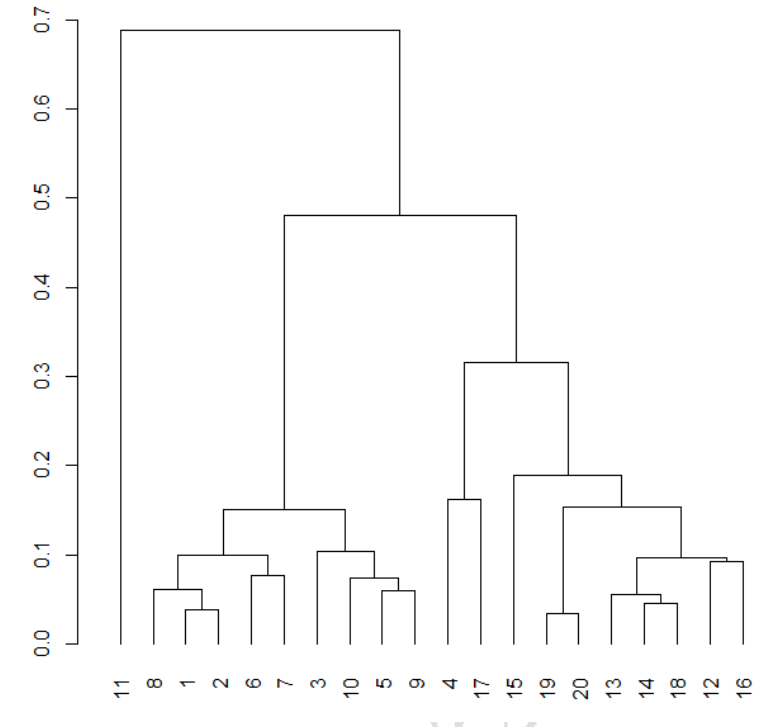
```
mydata<-read.table("clipboard",header = T)
lm.1<-lm(y~1,data=mydata)
lm.both=step(lm.1,scope = list(upper=~x1+x2+x3+x4,lower=~1),direction="both")
summary(lm.both)
lm1<-lm(y~x1+x2+x3,data = mydata)
summary(lm1)
lm2<-lm(y~x1+x2,data=mydata)
summary(lm2)
```

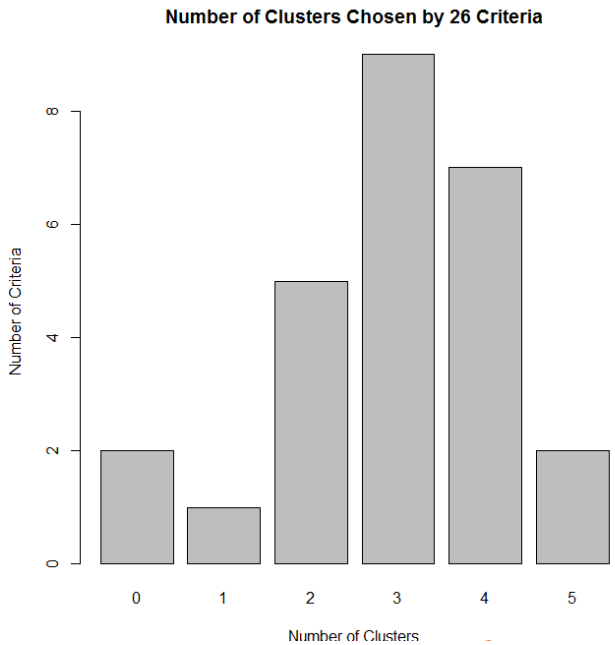
## 2 聚类分析

| a | t | c | g |
|---|---|---|---|
| 0.2973 | 0.1351 | 0.1712 | 0.3964 |
| 0.2703 | 0.1532 | 0.1622 | 0.4144 |
| 0.2703 | 0.0631 | 0.2162 | 0.4505 |
| 0.4234 | 0.2883 | 0.1081 | 0.1802 |
| 0.2342 | 0.1081 | 0.2342 | 0.4234 |
| 0.3514 | 0.1261 | 0.1261 | 0.3964 |
| 0.3514 | 0.1892 | 0.0991 | 0.3604 |
| 0.2793 | 0.1892 | 0.1622 | 0.3694 |
| 0.2072 | 0.1532 | 0.2072 | 0.4324 |
| 0.1818 | 0.1364 | 0.2727 | 0.4091 |
| 0.0690 | 0.0575 | 0.0575 | 0.8161 |
| 0.3273 | 0.5000 | 0.0273 | 0.1455 |
| 0.2545 | 0.5182 | 0.1000 | 0.1273 |
| 0.3000 | 0.5000 | 0.0818 | 0.1182 |
| 0.2909 | 0.6455 | 0.0000 | 0.0636 |
| 0.3636 | 0.4636 | 0.0818 | 0.0909 |
| 0.3545 | 0.2636 | 0.2455 | 0.1364 |
| 0.2909 | 0.5000 | 0.1182 | 0.0909 |
| 0.2182 | 0.5636 | 0.1455 | 0.0727 |

| | | | |
|---|---|---|---|
| 0. 2000 | 0. 5636 | 0. 1727 | 0. 0636 |

使用 nbclust 包中 26 个评判准则得到推荐聚类个数为 3
第一类：1，2， 3，5，6， 7， 8， 9， 10；
第二类：4， 12， 13， 14， 15， 16， 17， 18， 19， 20；
第三类：11




```
mydata<-read.table("clipboard",header = T)
hc<-hclust(dist(mydata),"ave")
plot(hc,hang = -1,cex=.8,main="Average Linkage Clustering")
dend<-as.dendrogram(hc)
```

```
library(NbClust) #引入做聚类的包
devAskNewPage(ask=TRUE)#请求绘制在一个新的画板上
nc<-NbClust(mydata,min.nc=2,max.nc=20,method="ave")
barplot(table(nc$Best.n[1,]),xlab="Number of Clusters",ylab="Number of Criteria",main="Number of Clusters Chosen by 26 Criteria")
```

## 3 判别分析

| x | y | z | n |
|---|---|---|---|
| 436.7 | 49.59 | 2.32 | 1 |
| 290.67 | 30.02 | 2.46 | 1 |
| 352.53 | 36.23 | 2.36 | 1 |
| 510.47 | 67.64 | 1.73 | 2 |
| 510.41 | 62.71 | 1.58 | 2 |
| 470.3 | 54.4 | 1.68 | 2 |
| 364.12 | 46.26 | 2.09 | 2 |

```
mydata<-read.table("clipboard",header = T)
library(e1071)
classifier<-naiveBayes(mydata[,1:3], as.factor(mydata[,4]))
pe<-data.frame(x=400.72,y=49.46,z=2.25)
predict(classifier, pe)
```

```
> pe<-data.frame(x=400.72,y=49.46,z=2.25)
> predict(classifier, pe)
[1] 1
Levels: 1 2
```

用原始数据建立贝叶斯判别模型，预测得到属于第一类即这个人是属于健康人

## 4 主成分分析

| 净产值利润率 | 固定资产利润率 | 总产值利润率 | 销售收入利润率 | 产品成本利润率 | 物耗利润率 | 人均利润率 | 流动资金利润率 |
|---|---|---|---|---|---|---|---|
| 40.4 | 24.7 | 7.2 | 6.1 | 8.3 | 8.7 | 2.442 | 20 |
| 25 | 12.7 | 11.2 | 11 | 12.9 | 20.2 | 3.542 | 9.1 |
| 13.2 | 3.3 | 3.9 | 4.3 | 4.4 | 5.5 | 0.578 | 3.6 |
| 22.3 | 6.7 | 5.6 | 3.7 | 6 | 7.4 | 0.176 | 7.3 |
| 34.3 | 11.8 | 7.1 | 7.1 | 8 | 8.9 | 1.726 | 27.5 |
| 35.6 | 12.5 | 16.4 | 16.7 | 22.8 | 29.3 | 3.017 | 26.6 |
| 22 | 7.8 | 9.9 | 10.2 | 12.6 | 17.6 | 0.847 | 10.6 |
| 48.4 | 13.4 | 10.9 | 9.9 | 10.9 | 13.9 | 1.772 | 17.8 |
| 40.6 | 19.1 | 19.8 | 19 | 29.7 | 39.6 | 2.449 | 35.8 |
| 24.8 | 8 | 9.8 | 8.9 | 11.9 | 16.2 | 0.789 | 13.7 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 12.5 | 9.7 | 4.2 | 4.2 | 4.6 | 6.5 | 0.874 | 3.9 |
| 1.8 | 0.6 | 0.7 | 0.7 | 0.8 | 1.1 | 0.056 | 1 |
| 32.3 | 13.9 | 9.4 | 8.3 | 9.8 | 13.3 | 2.126 | 17.1 |
| 38.5 | 9.1 | 11.3 | 9.5 | 12.2 | 16.4 | 1.327 | 11.6 |

（1）先建立 8 个主成分的模型，根据运行结果，以累积贡献率超过 90%为标准，可选择三个主成分

```
> pc8
Principal Components Analysis
Call: principal(r = mydata, nfactors = 8, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
                PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8 h2       u2 com
净产值利润率   0.80  0.42 -0.30 -0.31 -0.01  0.01  0.01  0.01  1  1.1e-16 2.2
固定资产利润率 0.73  0.61  0.07  0.17  0.24 -0.01  0.00  0.00  1 -4.4e-16 2.3
总产值利润率   0.96 -0.23 -0.03 -0.11  0.05  0.01 -0.03 -0.02  1  6.7e-16 1.2
销售收入利润率 0.95 -0.28  0.04 -0.05 -0.01 -0.08  0.00  0.00  1  1.2e-15 1.2
产品成本利润率 0.94 -0.32 -0.02  0.07  0.06  0.02  0.04 -0.01  1  1.3e-15 1.3
物耗利润率     0.92 -0.38  0.05  0.03  0.08  0.03 -0.02  0.02  1  1.6e-15 1.4
人均利润率     0.79  0.28  0.51 -0.06 -0.17  0.01  0.00  0.00  1  8.9e-16 2.1
流动资金利润率 0.88  0.16 -0.28  0.26 -0.23  0.00 -0.01  0.00  1  1.0e-15 1.6

                       PC1  PC2  PC3  PC4  PC5  PC6 PC7 PC8
SS loadings           6.14 1.04 0.44 0.22 0.15 0.01   0   0
Proportion Var        0.77 0.13 0.05 0.03 0.02 0.00   0   0
Cumulative Var        0.77 0.90 0.95 0.98 1.00 1.00   1   1
Proportion Explained  0.77 0.13 0.05 0.03 0.02 0.00   0   0
Cumulative Proportion 0.77 0.90 0.95 0.98 1.00 1.00   1   1

                PC1   PC2   PC3   h2     u2 com
净产值利润率   0.80  0.42 -0.30 0.90 0.0988 1.8
固定资产利润率 0.73  0.61  0.07 0.91 0.0885 2.0
总产值利润率   0.96 -0.23 -0.03 0.98 0.0151 1.1
销售收入利润率 0.95 -0.28  0.04 0.99 0.0095 1.2
产品成本利润率 0.94 -0.32 -0.02 0.99 0.0111 1.2
物耗利润率     0.92 -0.38  0.05 0.99 0.0097 1.3
人均利润率     0.79  0.28  0.51 0.97 0.0323 2.0
流动资金利润率 0.88  0.16 -0.28 0.88 0.1203 1.3

                       PC1  PC2  PC3
SS loadings           6.14 1.04 0.44
Proportion Var        0.77 0.13 0.05
Cumulative Var        0.77 0.90 0.95
Proportion Explained  0.81 0.14 0.06
Cumulative Proportion 0.81 0.94 1.00
```

（2）再建立 3 个主成分的模型，根据运行结果根据特征向量可以写出主成分表达式如第一主成分可写为如下，其它类似：

Pc1=0.13x1+0.12x2+0.16x3+0.16x4+0.15x5+0.15x6+0.13x7+0.14x8

```
> pc3$weights
          PC1        PC2         PC3
V1 0.1296341  0.4066272 -0.68340078
V2 0.1191514  0.5854579  0.15604476
V3 0.1570792 -0.2250519 -0.06042200
V4 0.1553045 -0.2730015  0.08159461
V5 0.1532166 -0.3098571 -0.05648063
V6 0.1497110 -0.3639183  0.11387221
V7 0.1291589  0.2724645  1.16707982
V8 0.1434917  0.1536334 -0.64334634
```

（3）可见，在第一主成分上得分最高的是企业 9，在第二主成分上得分最高的是企业 1，在第三主成分上得分最高的是企业 2。

```
              PC1           PC2         PC3
 [1,]  0.29534950  2.564339228  0.2719897
 [2,]  0.42980296 -0.008251151  2.7988195
 [3,] -1.14001610 -0.542432353  0.1619599
 [4,] -0.88349335 -0.155489023 -0.9265355
 [5,]  0.02696288  1.020274591 -0.9742650
 [6,]  1.39583929 -0.849798458  0.6083021
 [7,] -0.11169174 -1.029946655 -0.2378226
 [8,]  0.40818194  0.854897865 -0.9755337
 [9,]  2.10875248 -1.000761635 -0.6316778
[10,] -0.12192347 -0.735322862 -0.6711424
[11,] -0.95673735  0.077931163  0.6719126
[12,] -1.75926321 -0.681542845  0.2442254
[13,]  0.16113589  0.701299526  0.3105301
[14,]  0.14710027 -0.215197392 -0.6507624
> |
```

代码：

```
library(psych)
mydata<-read.table("clipboard",header = T)
pc8<-principal(mydata,nfactors = 8,rotate = "none")
pc8
pc3<-principal(mydata,nfactors = 3,rotate = "none",scores = TRUE)
pc3
pc3$scores
pc$weights
```

**第 5 题**

| x1 | x2 |
|---|---|
| 10.2 | 9.5 |
| 9.6 | 9.8 |
| 9.2 | 8.8 |
| 10.6 | 10.1 |
| 9.9 | 10.3 |
| 9.1 | 9.3 |
| 10.6 | 10.5 |
| 10 | 10 |
| 11.2 | 10.6 |
| 10.7 | 10.2 |
| 10.6 | 9.8 |

```
mydata<-read.table("clipboard",header = T)
attach(mydata)
wilcox.test(x1,x2)
```

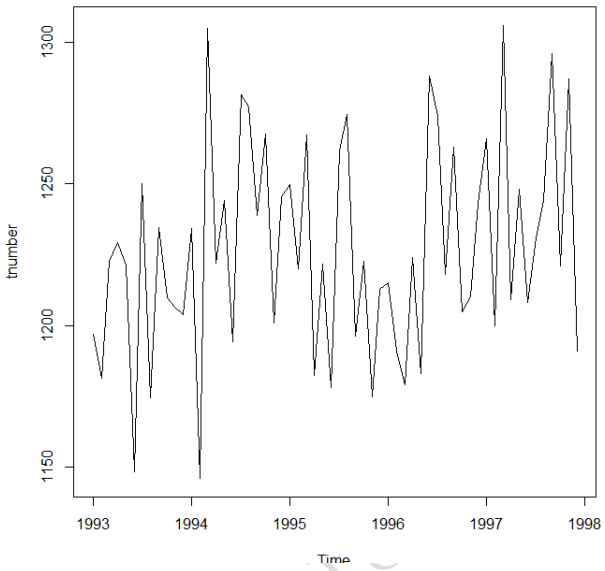
运行结果为

```
> wilcox.test(x1,x2)

        Wilcoxon rank sum test with continuity correction

data:  x1 and x2
W = 76.5, p-value = 0.307
alternative hypothesis: true location shift is not equal to 0
```

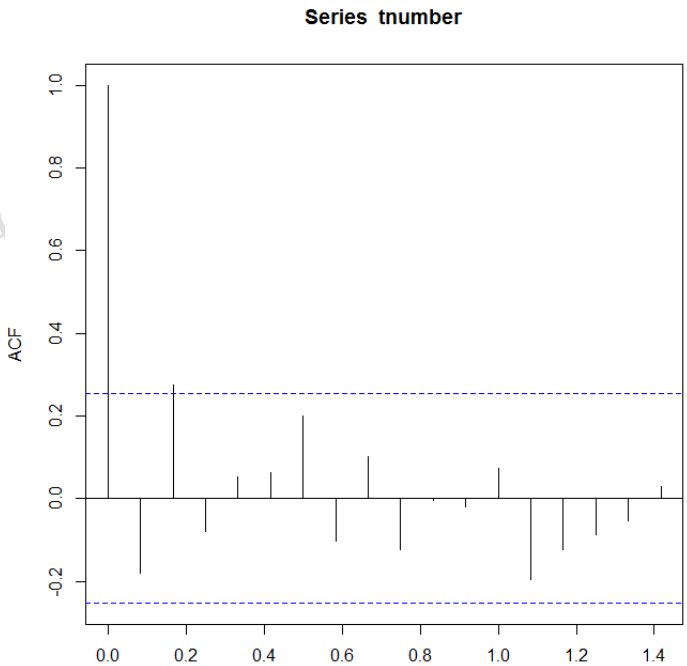
检验其显著性为 0.307，大于 0.05，故接受原假设，认为这两组方法没有显著性差异。

## 第 6 题

（1）由时序图可知，此车站列车运行数量数据在一个常数值附近随机波动，而且波动范围有界，无明显趋势及周期特征，基本可以视序列为平稳序列。




（2）由此自相关图可看出，自相关系数很快的衰减向 0，且基本控制在 2 倍范围内，可以认为该序列为平稳序列。




（3）由于统计量 P 值均大于 0.05，则认为在 0.05 的显著水平下，无法拒绝原假设，即不能显著拒绝序列为纯随机序列的假定，因而认为此车站列车运行数量为纯随机波动序列，各序列之间没有任何行相关关系，即为无记忆序列，也就是说，

该车站列车运行数量前后两年并无大的联系，也就是实说，我们很难根据历史信息预测未来年份此车站列车运行数量，故，该平稳序列不值得继续分析下去，对该序列分析到此结束。

```
> Box.test(tnumber, type="Ljung-Box",lag=6) #

        Box-Ljung test

data:  tnumber
X-squared = 10.504, df = 6, p-value = 0.105

> Box.test(tnumber, type="Ljung-Box",lag=12)

        Box-Ljung test

data:  tnumber
X-squared = 13.52, df = 12, p-value = 0.3324
```

代码：

```
number<-c(1196.8,1181.3,1222.6,1229.3,1221.5,1148.4,1250.2 ,1174.4,1234.5,1209.7,
          1206.5,1204,1234.1,1146,1304.9,  1221.9,  1244.1,  1194.4  ,1281.5  ,1277.3,
          1238.9,1267.5,1200.9,1245.5,1249.9,  1220.1,  1267.4,1182.3,  1221.7,  1178.1,
          1261.6,1274.5,1196.4,1222.6,1174.7,1212.6,  1215,  1191,  1179,  1224,
          1183,1288,1274,1218,1263,1205,1210,1243,  1266,1200,
          1306,1209,1248,1208,1231,1244,1296,1221,1287,  1191
)
tnumber<-ts(number,start = c(1993,1),frequency = 12)
plot(tnumber)
acf(tnumber)
Box.test(tnumber, type="Ljung-Box",lag=6) #纯随机性检验
Box.test(tnumber, type="Ljung-Box",lag=12)
```