毕业论文 Rmarkdown 代码

吴欣宜 161070073

目录

1 聚类分析

1.1 读入数据

```
d2<-read.csv("22.csv")
d2
```

```
##    X...Name Revenue
## 1     wanda   88.78
## 2      gddd   62.30
## 3      shlh   51.00
## 4     nfxgx   46.26
## 5      zysz   44.73
```

```
## 6      zyxm   33.82
## 7    gzjyzj   31.43
## 8      zjhd   27.02
## 9      xflh   25.11
## 10     hxlh   23.47
```

1.2 聚类分析

参考文档解释“我们分析一下结果，第一行表示各个类别下数据点的数量分别是 96、21 和 33 个。然后是聚类的均值，即聚类的中心点。然后是聚类向量，表明每个数据点所属的类别。Within cluster sum of squares by cluster 表示每个簇内部的距离平方和，表示该簇的紧密程度。between_SS / total_SS 这一项表示组间距离的平方和占整体距离平方和的结果。一般的，组内距离要求尽可能小，组间距离尽可能大，因此这个值越接近 1 越好。最后的 Available components 表示运行结果返回的对象包含的组成部分。可以使用 km$cluster 形式打印出来查看结果。”

```
km1 <- kmeans(d2[2],3)
km2 <- kmeans(d2[2],4)#三个和四个分别聚类比较结果
km1
```

```
## K-means clustering with 3 clusters of sizes 4, 1, 5
##
## Cluster means:
##   Revenue
## 1 51.0725
## 2 88.7800
## 3 28.1700
##
## Clustering vector:
##  [1] 2 1 1 1 1 3 3 3 3 3
##
## Within cluster sum of squares by cluster:
## [1] 189.4495   0.0000  75.3262
##  (between_SS / total_SS =  92.9 %)
```

```
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```
km2#四个的情况bss/tss更大，更优
```

```
## K-means clustering with 4 clusters of sizes 2, 4, 1, 3
##
## Cluster means:
##   Revenue
## 1 32.6250
## 2 51.0725
## 3 88.7800
## 4 25.2000
##
## Clustering vector:
##  [1] 3 2 2 2 2 1 1 4 4 4
##
## Within cluster sum of squares by cluster:
## [1]   2.85605 189.44948   0.00000   6.31340
##  (between_SS / total_SS =  94.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

2 二、集中度实证检验

2.1 读入数据

```
data0<-read.csv("Data.csv",header=T)
data0
```

```
##    X...Year    CR4 C.t.1.      C    Au        I      E   T    CB1     CB2
## 1      2010 0.4577 0.4680 0.6391  2.37 16900.50 0.0000  14 0.0163  2.7492
## 2      2011 0.4233 0.4577 0.2893  3.45 21426.90 0.0541  61 0.0141  3.3628
## 3      2012 0.4258 0.4233 0.2629  4.62 24126.70 0.0256  98 0.0126  4.1408
## 4      2013 0.3872 0.4258 0.3143  6.12 26467.00 0.0500 140 0.0120  5.1831
## 5      2014 0.3783 0.3872 0.3615  8.30 28843.85 0.0714 203 0.0126  6.5864
## 6      2015 0.3712 0.3783 0.4869 12.60 31194.83 0.0222 278 0.0139  9.5802
## 7      2016 0.3404 0.3712 0.1183 13.72 33616.25 0.0000 394 0.0120 10.7137
## 8      2017 0.3459 0.3404 0.1345 16.20 36396.19 0.0435 513 0.0110 11.6481
## 9      2018 0.3608 0.3459 0.0906 17.16 39250.84 0.0000 609 0.0101 12.7033
## 10     2019 0.3864 0.3608 0.0540 17.27 30733.00 0.0417 717 0.0092 12.8532
##       CR8
## 1  0.6824
## 2  0.6404
## 3  0.6566
## 4  0.6105
## 5  0.5927
## 6  0.5972
## 7  0.5641
## 8  0.5716
## 9  0.5736
## 10 0.5996
```

2.2 相关性分析

绝对值 <0.5 剔除, 因此 E 剔除

```
res<-cor(data0)
round(res,3)[2,3:10]
```

```
## C.t.1.      C     Au      I      E      T    CB1    CB2
##  0.884  0.630 -0.823 -0.924 -0.004 -0.690  0.658 -0.821
```

2.3 简单 OLS 回归

判定显著性

```
f1<-lm(CR4~C.t.1.+C+Au+I+T+CB1+CB2,data=data0)
summary(f1)
```

```
##
## Call:
## lm(formula = CR4 ~ C.t.1. + C + Au + I + T + CB1 + CB2, data = data0)
##
## Residuals:
##          1          2          3          4          5          6
## -2.401e-03 -1.998e-03  1.862e-02 -8.047e-03 -1.114e-02  9.377e-03
##          7          8          9         10
## -6.082e-03 -5.159e-03  6.841e-03 -1.765e-05
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.788e-01  4.113e-01   0.921    0.454
## C.t.1.      -6.912e-03  8.968e-01  -0.008    0.995
## C            5.150e-02  9.204e-02   0.560    0.632
## Au          -6.888e-03  3.476e-02  -0.198    0.861
## I           -7.000e-07  4.595e-06  -0.152    0.893
## T            3.440e-04  2.432e-04   1.414    0.293
## CB1          6.706e+00  1.261e+01   0.532    0.648
## CB2         -1.249e-02  4.116e-02  -0.303    0.790
##
## Residual standard error: 0.01928 on 2 degrees of freedom
```

```
## Multiple R-squared:  0.9414, Adjusted R-squared:  0.7363
## F-statistic:  4.59 on 7 and 2 DF,  p-value: 0.1905
```

2.4 VIF 检验多重共线性

“在进行线性回归分析时，很容易出现自变量共线性问题，通常情况下 VIF 值大于 10 说明严重共线，VIF 大于 5 则说明有共线性问题；当出现共线性问题时，可能导致回归系数的符号与实际情况完全相反，本应该显著的自变量不显著，本不显著的自变量却呈现出显著性；共线性问题会导致数据研究出来严重偏差甚至完全相反的结论，因而需要解决此问题。”

```
library(car)
```

```
## Loading required package: carData
```

```
vif(f1,digits=3)
```

```
##      C.t.1.           C          Au           I           T         CB1
##   39.951517    7.154352 1010.249255   24.003391   85.701400   16.349365
##         CB2
##  650.341509
```

```
#超过10都存在多重共线性，需要剔除 Au和ECn都超过200
plot(data0[,c(3:6,8:10)])
```

[](<95><9a><96><87><81>_files/figure-latex/plot-1.pdf)

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
corrplot::corrplot(cor(data0[,c(3:6,8:10)]),diag = FALSE)
```

[](<95><9a><96><87><81>_files/figure-latex/corrplot-1.pdf)

```
#色谱图
```

2.5 多重共线性的解决—岭回归

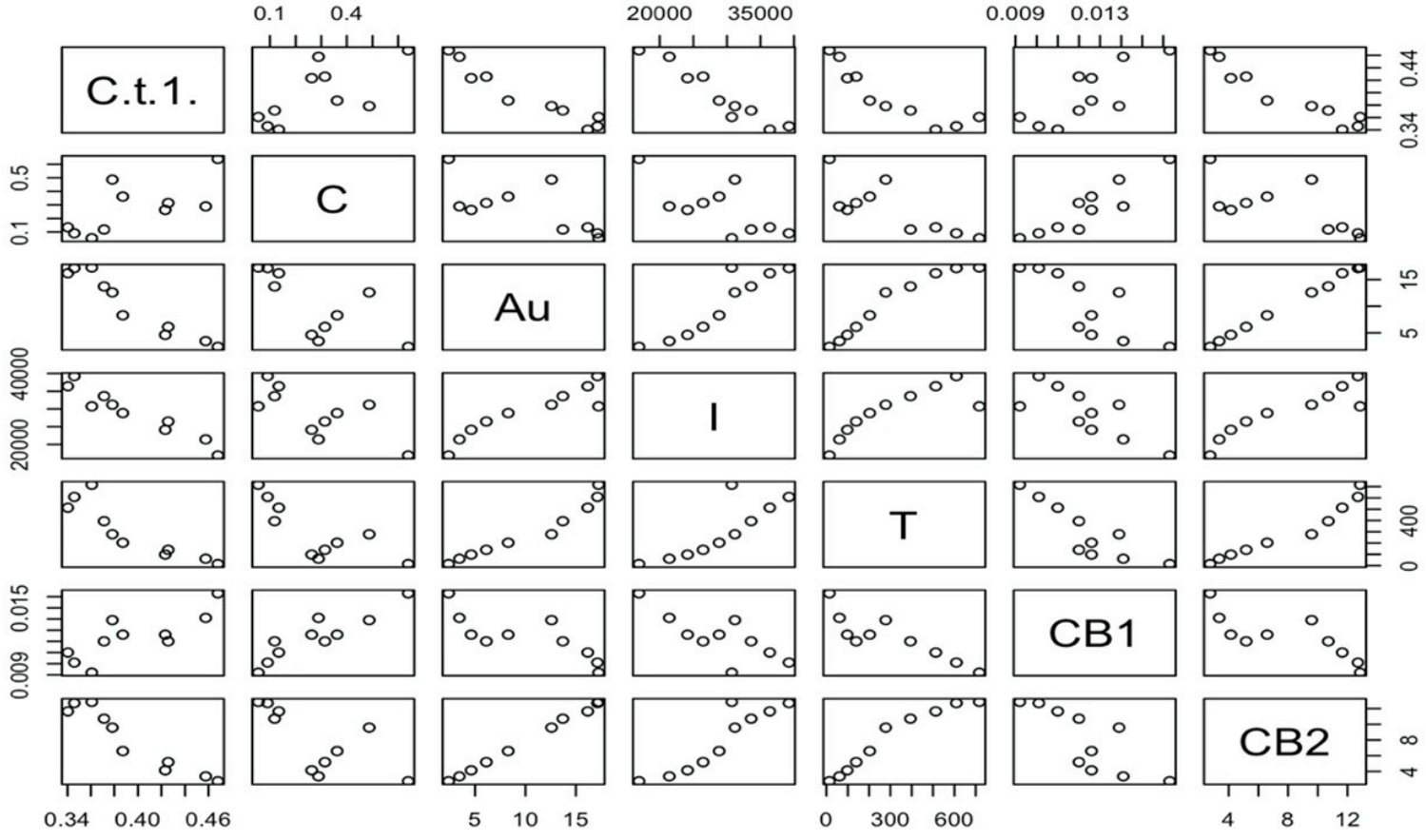
```
library(ridge)#岭回归
```

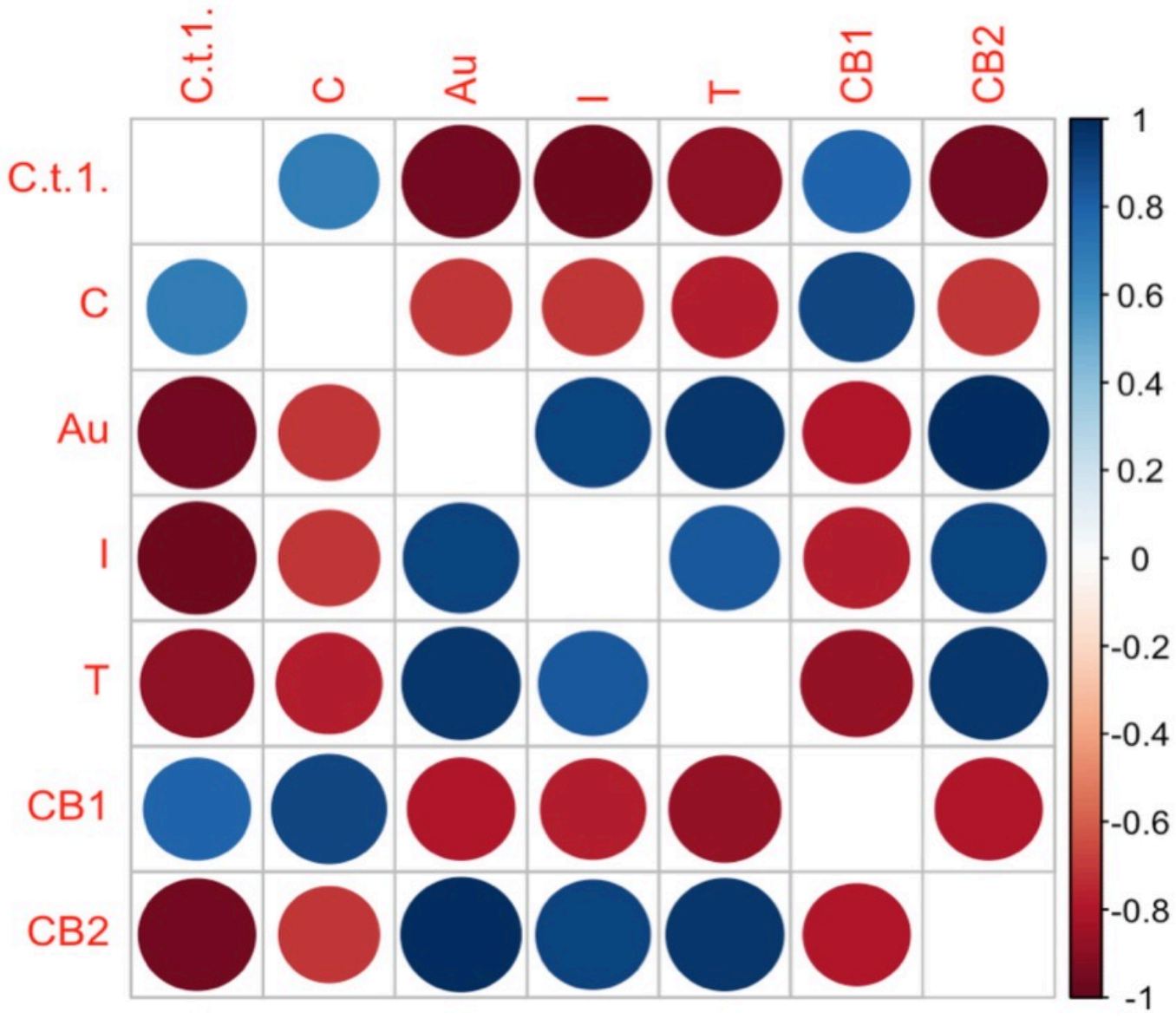
```
## Warning: package 'ridge' was built under R version 3.5.2
```

```
mod <- linearRidge(CR4~C.t.1.+C+Au+I+T+CB1+CB2, data = data0)
summary(mod)#Ct.1,Au,Line显著，其他不显著
```

```
##
## Call:
## linearRidge(formula = CR4 ~ C.t.1. + C + Au + I + T + CB1 + CB2,
##     data = data0)
##
##
## Coefficients:
##               Estimate Scaled estimate Std. Error (scaled)
## (Intercept)  3.745e-01              NA                   NA
## C.t.1.       2.104e-01       2.859e-02            1.009e-02
## C            1.671e-02       9.364e-03            1.373e-02
## Au          -8.041e-04      -1.418e-02            6.561e-03
## I           -2.069e-06      -4.254e-02            1.234e-02
## T            1.804e-05       1.324e-02            1.031e-02
## CB1         -1.855e-01      -1.147e-03            1.275e-02
## CB2         -1.243e-03      -1.485e-02            7.372e-03
##             t value (scaled) Pr(>|t|)
## (Intercept)               NA       NA
## C.t.1.                 2.833 0.004608 **
## C                      0.682 0.495129
## Au                     2.161 0.030716 *
## I                      3.447 0.000567 ***
## T                      1.284 0.198998
```

```
## CB1                    0.090 0.928338
## CB2                    2.015 0.043928 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Ridge parameter: 0.3007845, chosen automatically, computed using 1 PCs
##
## Degrees of freedom: model 2.319 , variance 1.548 , residual 3.091
```


C.t.1.
C
Au
I
T
CB1
CB2
0.1
0.4
20000
35000
0.009
0.013
0.34
0.44
0.1
0.5
5
15
20000
40000
0
400
0.009
0.015
4
8
0.34
0.40
0.46
5
10
15
0
300
600
4
8
12


C.t.1.
C
Au
I
T
CB1
CB2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1