**云南大学数学与统计学院**

**实验报告**

**实验课名称： 应用多元统计分析实验**

**指导教师： 李会琼**

**专业（年级）： 统计学2021级**

**学生姓名： 枫叶 学号:**

**实验名称： 实验9**

**实验成绩：**

**《应用多元统计分析实验》实验报告 9**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 实验名称 |  | | 实验成绩 |  | |
| 学号 |  | | 姓名 | 枫叶 | |
| 实验时间 | 2024年5月28日 | 实验地点 | 格物楼3508 | 指导教师 | 李会琼 |
| 1. **实验目的**   学习使用R软件进行距离判别   1. **实验要求**   1. 对所使用的方法与所得到的结果进行适当的文字描述。  2. 在实验结果的相应部分附上完整的代码与适当的注释。  3. 采用一定的可视化方法体现出对应计算结果。   1. **实验内容**   例题6.4.1 习题6-9 习题6-10 习题6-11   1. **实验软件**   R语言   1. **实验结果**  加载包 library(readxl) library(MASS) library(dplyr) library(NbClust) 第一题 data <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计表6.7数据.xlsx") %>%  select(-1) %>%  apply(2,scale) D <- dist(data) model <- hclust(D,method = "average") plot(model)  #R2(平替) NbClust(as.matrix(data),distance = "euclidean",method = "average",min.nc=2,max.nc=10,index = "tracew")  ## $All.index ## 2 3 4 5 6 7 8 9 10  ## 49.8720 44.8197 27.3054 15.9760 10.9695 8.5819 6.9176 4.7856 3.9622  ##  ## $Best.nc ## Number\_clusters Value\_Index  ## 5.0000 6.3229  ##  ## $Best.partition ## [1] 1 2 3 3 3 2 2 4 5 2 2 4 4 4 2 3  #半偏R2(平替) a <- NbClust(as.matrix(data),distance = "euclidean",method = "average",min.nc=2,max.nc=10,index = "tracew")$All.index -diff(a)  ## 3 4 5 6 7 8 9 10  ## 5.0523 17.5143 11.3294 5.0065 2.3876 1.6643 2.1320 0.8234  #伪F NbClust(as.matrix(data),distance = "euclidean",method = "average",min.nc=2,max.nc=10,index = "ch")  ## $All.index ## 2 3 4 5 6 7 8 9 10  ## 11.2647 6.5523 9.1842 12.7420 14.4092 14.2308 13.7261 15.5807 14.4764  ##  ## $Best.nc ## Number\_clusters Value\_Index  ## 9.0000 15.5807  ##  ## $Best.partition ## [1] 1 2 3 3 3 4 5 6 7 8 8 9 9 9 2 3  #伪t2 NbClust(as.matrix(data),distance = "euclidean",method = "average",min.nc=2,max.nc=10,index = "pseudot2")  ## $All.index ## 2 3 4 5 6 7 8 9 10  ## 0.0000 7.6970 8.1992 3.6933 0.5155 0.0000 2.6223 0.0000 0.4998  ##  ## $All.CriticalValues ## 2 3 4 5 6 7 8 9 10  ## 0.0000 15.6060 14.8210 16.7852 55.4763 0.0000 55.4763 0.0000 55.4763  ##  ## $Best.nc ## Number\_clusters Value\_Index  ## 2 0  ##  ## $Best.partition ## [1] 1 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2  教材中基于,半偏,伪和伪统计量进行了聚类个数的判断，为方便起见，这里使用Tracew统计量作为的替代，其效果是一致的。Tracew统计量即类内平方和，可以看到在类别数大于4时其变化幅度不大，而在4类及之后变化幅度剧增，那么可以考虑分为4个类；考虑将Tracew统计量差分，取负后则可以得到与半偏效果一致的统计量，可以看到变化最大的是4和5，故可以考虑分为4类或5类；若将类别数限制在6以内(过多类别会导致过拟合)，则伪F统计量最大的是5和2，可以考虑分为5类或2类；结合临界值来看，分为2、4或5类是较好的，下面给出4类和5类的结果  plot(model) rect.hclust(model,k = 4)    plot(model) rect.hclust(model,k = 5)    其中分为5类的结果与教材是一致的，即河北、河南、山西、内蒙古为一类，黑龙江、安徽、福建、江西为一类，天津、山东、江苏、浙江、辽宁、吉林为一类，北京和上海各自为一类 第二题 data <- c(1,2,5,7,10) D <- dist(data) plot(hclust(D,"ward.D"))    ,, , , ,各类别数情况下的总离差平方和如下  W\_k <- function(x,group){  data <- data.frame(x,group) %>%  group\_by(group) %>%  mutate(mean=mean(x),  dev1=(x-mean)^2) %>%  summarise(dev=sum(dev1))  return(sum(data$dev))  } W\_k(x=data,group = c(1,2,3,4,5))#k=5  ## [1] 0  W\_k(x=data,group = c(1,1,2,3,4))#k=4  ## [1] 0.5  W\_k(x=data,group = c(1,1,2,2,3))#k=3  ## [1] 2.5  W\_k(x=data,group = c(1,1,2,2,2))#k=2  ## [1] 13.16667  W\_k(x=data,group = c(1,1,1,1,1))#k=1  ## [1] 54 第二题Q聚类 data <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计习题6-10数据.xlsx") %>%  apply(2,scale) D <- dist(data) plot(hclust(D,"single"))    plot(hclust(D,"complete"))    plot(hclust(D,"median"))    plot(hclust(D,"centroid"))    plot(hclust(D,"average"))    plot(hclust(D,"mcquitty"))    plot(hclust(D,"ward.D"))    对样本进行聚类分析的结果如上述谱系图所示，可以看到所有方法都倾向于将1和2分为一类，后续也倾向于将4和6,3和5分为一类或是逐步引入4,6,3,5 R聚类直接转置 data <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计习题6-10数据.xlsx") %>%  t() D <- dist(data) plot(hclust(D,"single"))    plot(hclust(D,"complete"))    plot(hclust(D,"median"))    plot(hclust(D,"centroid"))    plot(hclust(D,"average"))    plot(hclust(D,"mcquitty"))    plot(hclust(D,"ward.D"))    对数据阵直接转置，此时不宜再使用标准化，使用系统聚类法对变量进行聚类，聚类分析结果如上述谱系图所示，可以看到所有方法的结果都是一致的 基于相关系数转化距离矩阵 为简便起见，用相关系数度量变量间相似关系，并定义距离为  R <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计习题6-10数据.xlsx") %>%  cor() D <- 1-abs(R) %>% as.dist() plot(hclust(D,"single"))    plot(hclust(D,"complete"))    plot(hclust(D,"median"))    plot(hclust(D,"centroid"))    plot(hclust(D,"average"))    plot(hclust(D,"mcquitty"))    plot(hclust(D,"ward.D"))    基于各种方法进行聚类的谱系图如上所示，可以注意到所有方法都倾向于将X3和X4分为一类，且除了WARD法和最长距离法之外，所有方法都倾向于将X3,X4和X7分为一类 第四题 data <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计习题5-9数据.xlsx") %>%  select(-2) %>%  mutate(across(1,factor,labels=c(1,2)))  ## Warning: There was 1 warning in `mutate()`. ## ℹ In argument: `across(1, factor, labels = c(1, 2))`. ## Caused by warning: ## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0. ## Supply arguments directly to `.fns` through an anonymous function instead. ##  ## # Previously ## across(a:b, mean, na.rm = TRUE) ##  ## # Now ## across(a:b, \(x) mean(x, na.rm = TRUE))  data[-1]%>%  apply(2,scale)  ## Cu Ag Bi ## [1,] -0.08688559 -1.7223269 -0.32092614 ## [2,] 0.54093287 -0.8391804 -0.13214606 ## [3,] 1.81618911 -1.0532765 -0.13214606 ## [4,] -0.53813011 -1.0532765 -0.92502239 ## [5,] 1.79656978 0.8200646 -0.92502239 ## [6,] 0.14854633 1.8370213 1.00053443 ## [7,] 0.14854633 0.4186344 -2.09545889 ## [8,] -0.73432337 1.1679709 0.09439004 ## [9,] -0.91089731 0.6862546 0.09439004 ## [10,] -0.57736876 0.5256825 0.24541410 ## [11,] -1.30328385 -0.1701300 -0.13214606 ## [12,] -1.34252250 -0.3842261 -0.13214606 ## [13,] 0.73712613 -0.6518463 1.68014272 ## [14,] 0.30550095 0.4186344 1.68014272  D <- dist(data) plot(hclust(D,"single"))    plot(hclust(D,"complete"))    plot(hclust(D,"median"))    plot(hclust(D,"centroid"))    plot(hclust(D,"average"))    plot(hclust(D,"mcquitty"))    plot(hclust(D,"ward.D"))    可以看到在两个类别的假定下，大部分聚类方法都将样本完美地作了分类，仅有中间距离法和重心法表现不佳 | | | | | |