多元统计分析——R 陪同

实 验 指 导 书

学生专业　　应用统计学

任课教师　　董志清

目录

实验题目 0 金融数据下载预处理

实验目的

通过上机试验，熟悉 R 软件基本操作过程，熟悉用 R 语言对金融数据的下载与整理。

实验工具

计算机，Windows，R 软件

实验内容

1、下载股票系统数据并合并数据为一个 excel 表格；

2、计算数据矩阵的均值、协方差和相关系数，并进行皮尔逊相关性检验[杨虎]p12；

3、进行简单描述性统计分析；直观图（轮廓图、星座图、脸谱图、调和曲线图、均值条图、箱图）；

实验指导

1、下载股票系统数据并合并数据为一个 excel 表格

```
setwd("G:\\证券数据")
a=list.files("G:\\证券数据")          #读取所有文件并保存
n=length(a)
zd=read.csv(file = a[1],header = F,sep = ",")     #读取第一个文件
for(i in 2:n)
{
  new.data=read.csv(file = a[i],header = F,sep = ",")
  new.data<-new.data[3,]              #去除无用表格头
  zd=rbind(zd,new.data)               #按行合并
}
p<-zd[-c(1,2,4),]
p$V1<-as.Date(p$V1,"%Y/%m/%d")       #将因子型时间格式转换为标准时间格式
names(p)<-c("日期","开盘","最高","最低","收盘","成交量","成交额")
write.csv(p,file = "数据合并.csv",row.names = F)       #保存合并数据
```

2、计算数据矩阵的均值、协方差和相关系数，并进行皮尔逊相关性检验[1]p12；

```
x1<-c(1,2,3,3,1,2,1,2,0) ##录入数据##
x2<-c(1,2,3,3,1,2,2,3,1)
x3<-c(1,2,0,0,1,2,2,0,1)
x4<-c(0,1,2,1,2,0,2,3,1)
m1<-matrix(x1,nrow = 3,ncol = 3,byrow = F) ##按列合并为矩阵##
m2<-matrix(x2,nrow = 3,ncol = 3,byrow = F)
m3<-matrix(x3,nrow = 3,ncol = 3,byrow = F)
m4<-matrix(x4,nrow = 3,ncol = 3,byrow = F)
colMeans(m1);colMeans(m2);colMeans(m3);colMeans(m4);
cov(m1);cov(m2);cov(m3);cov(m4);
```

```
cor(m1);cor(m2);cor(m3);cor(m4);
install.packages("psych")
library(psych)
cor.test(m1[,1],m1[,2],method = "pearson")
cor.test(m1[,1],m1[,3],method = "pearson")
cor.test(m1[,2],m1[,3],method = "pearson")
setwd("G:\\证券数据")
mydata<-read.csv(file = "数据合并.csv",header= T)     #读取证券文件
mydata<-as.data.frame(mydata)                          #转换为数据框格式
mydata$开盘<-as.numeric(as.character(mydata$开盘)) #把因子型数据转换为数值型数据
mydata$最高<-as.numeric(as.character(mydata$最高))
mydata$最低<-as.numeric(as.character(mydata$最低))
mydata$收盘<-as.numeric(as.character(mydata$收盘))
mydata$成交量<-as.numeric(as.character(mydata$成交量))
mydata$成交额<-as.numeric(as.character(mydata$成交额))
apply(mydata[,c(2:7)],2,mean)                  #求均值
col(mydata[,c(2:7)])                           #求相关系数
cor(mydata[,c(2:7)])                           #求协方差系数
cor.test(mydata[,3],mydata[,4],method = "pearson")   #对最高和最低价进行皮尔逊检验
cor.test(mydata[,3],mydata[,5],method = "pearson")   #对最高价和收盘价进行皮尔逊检验
cor.test(mydata[,4],mydata[,5],method = "pearson")   #对最低价和收盘价进行皮尔逊检验
```

3、进行简单描述性统计分析；直观图（轮廓图、星座图、脸谱图、调和曲线图、均值条图、箱图）；

```
city <- c("北京","上海","陕西","甘肃")
肉食及制品 <- c(536.51,678.92,237.38,253.41)
住房 <- c(227.78,365.07,174.48,156.13)
医疗保健 <- c(147.76,112.82,119.78,102.96)
交通与通讯 <- c(235.99,301.46,141.07,108.13)
文娱用品及服务 <- c(510.78,465.88,245.57,212.2)
data1 <- data.frame(city,肉食及制品,住房,医疗保健,交通与通讯,文娱用品及服务)
data2 <- as.matrix(data1)
barplot(data2[,c(2:5)])
head(data3)
summary(data2)
cor.test(data1$肉食及制品,data1$医疗保健,method = "pearson")##皮尔逊检验
b <- table(data1$肉食及制品,data1$医疗保健)
summary(b)
barplot(b)
pie(b)
outline <- function(x, txt = TRUE){
   if (is.data.frame(x) == TRUE)
      x <- as.matrix(x)
   m <- nrow(x); n <- ncol(x)
   plot(c(1,n), c(min(x),max(x)), type = "n",
```

```
            main = "The outline graph of Data",
            xlab = "Number", ylab = "Value")
    for(i in 1:m){
        lines(x[i,], col=i)
        if (txt == TRUE){
            k <- dimnames(x)[[1]]
            text(1+(i-1)%%n, x[i,1+(i-1)%%n], k)
        }
    }
}
outline(data1)
stars(data1[,c(2:5)])
boxplot(data1[,c(2:5)])
install.packages("aplpack")
library(aplpack)
faces(b)
hist(b)
```

实验题目 1 多元数据简单统计分析

实验目的

通过实验，掌握 R 语言统计分析中的一些基本运算技巧与分析方法，进一步加深对 R 语言统计分析的理解。

实验工具

计算机，Windows，R 软件

实验内容

1、计算数据矩阵的均值、协方差和相关系数，并进行皮尔逊相关性检验；

2、进行简单描述性统计分析；直观图（轮廓图、星座图、脸谱图、调和曲线图、均值条图、箱图）；

3、吴喜之 P222 实践任选 2 例验证。

实验指导

1、 计算数据矩阵的均值、协方差和相关系数，并进行皮尔逊相关性检验

```
data=as.matrix(mtcars)                  #使用 mtcars 数据集,并将其转为矩阵
cov(data)                               #返回矩阵内各变量之间的协方差
cor(data)                               #返回矩阵内各变量之间的相关系数
library(psych)
corr.test(data,method = "pearson")   #进行各变量之间皮尔逊相关性检验
```

2、进行简单描述性统计分析；直观图（轮廓图、星座图、脸谱图、调和曲线图、均值条图、箱图）

```
summary(mtcars)
#箱线图
boxplot(mtcars[,c(3,4)],col=c("lightblue","pink"))
boxplot(mtcars[,-c(3,4)],col =c(11:19) )
#均值条形图
barplot(apply(mtcars, 1, mean),col = c(1:nrow(mtcars)))   #按行计算均值条形图
barplot(apply(mtcars, 2, mean),col = c(1:ncol(mtcars)))    #按列计算均值条形图
#调和曲线图
msa.andrews(mtcars)
#脸谱图
library(aplpack)
faces(mtcars[c(1:9),])
#星象图
stars(mtcars[c(1:6),],col.stars = c(1:6))
stars(mtcars[c(1:6),],draw.segments = T)
#轮廓图
library(lattice)
parallelplot(mtcars)
```

3、吴喜之 P222 实践任选 2 例验证

```
#实践 1 关于抽样数据
x=1:100                                  #x 从 1 到 100 的等差数列，公差为 1
sample(x,20,replace = FALSE)             #不放回抽取 20 个样本，默认不放回
sample(x,20,replace = TRUE)              #有放回抽取 20 个样本
set.seed(0)   #设置随机种子
z=sample(1:200000,10000)
y=c(1,3,7,3,4,2)
z[y]            #以 Y 为下标的 z 的值
(z=sample(x,100,replace = TRUE))         #从 x 有放回抽样 100 个并显示
(z1=unique(z))                           #从 z 中返回不相等的数赋值给 z1 并显示
length(z1)                               #z 中不同元素的个数
apropos("len")          #apropos("len")可以帮你找到所有包含"len"字符的所有函数或数据
xz=setdiff(x,z)         #x 和 z 的集合差
sort(union(xz,z))        #合并 xz 和 z 后，排序
setequal(union(xz,z),x)             #对比 xz 和 z 的并，与 x 是否一致
(w=intersect(1:10,7:50))          #两个数据的交
sample(1:100,20,prob=1:100)   #从 1：100 中不等概率随机抽样

#实践 2 一些简单计算
pi*10^2
pi*(1:10)^-2.3
(x=pi*10^2)
pi^(1:5)
print(x,digit=20)   #输出 x 的 20 位数字

#实践 3 关于 R 对象的类型
class(pi);class("you are the apple of my eye");class(c(1:5))
class(mtcars)
names(mtcars)        #显示数据框中的变量名
summary(mtcars)
head(mtcars);tail(mtcars)
str(mtcars)
row.names(mtcars)
attributes(mtcars)     #数据的一些属性
class(mpg~cyl)
plot(mpg~cyl,mtcars)

#实践 4 回归-简单自变量为定量定性变量
ncol(cars);nrow(cars)          #cars 的行列数
dim(cars)                       #维数
lm(dist~speed,data=cars)     #以 dist 因变量，以 speed 自变量做 OLS
cars$qspeed=cut(cars$speed,breaks=quantile(cars$speed),include.lowest=TRUE)
#增加定性变量 qspeed,四分位点为分割点
```

```
names(cars)                    #这时，数据里多了一个变量 qspeed
cars[3]
table(cars[3])
is.factor(cars$qspeed)
plot(dist~qspeed,data=cars)#点出箱型图
#拟合线性模型（简单最小二乘回归）
(a=lm(dist~qspeed,data=cars))
summary(a)                     #回归结果（包括一些检验）

#实践 5  简单样本描述统计量(使用 mtcars 数据集)
x=mtcars
summary(x)      #查看 x 所有变量的一些数据特征
attach(x)        #以 mpg 变量为例
min(mpg);max(mpg)
range(mpg);median(mpg);mean(mpg);sd(mpg);sqrt(var(mpg))
rank(mpg)       #返回每个数值对应的秩
order(mpg)       #返回各个数值按升序的下标
order(mpg,decreasing = T)       #返回各个数值按降列的下标
sort(mpg)                       #升序排列 mpg
sort(mpg,decreasing = T)        #降序排列 mpg
sum(mpg);length(mpg)
round(mpg)                      #四舍五入 mpg 中的数据
round(1.225,2);round(1.225,1)
fivenum(mpg)                   #返回最小值、下四分位数、中位数、上四分位数、最大值
quantile(mpg)                  #返回五个分位点
quantile(mpg,c(0,1/3,2/3,1)) #返回四个分位点
mad(mpg)                       #返回 mpg 的绝对中位差函数
cummin(mpg)                    #求累计最小值，从左往右，分别返回当前位置以前的最小值
cummax(mpg);cumprod(mpg)     #求累计最大值和累计乘积
cor(mpg,mtcars$hp)              #返回 mpg 与 hp 的相关系数

#实践 6  简单图形
x=rnorm(380)
hist(x,col = "lightblue");rug(x)     #在直方图下面标出实际数据的位置
stem(x)                         #茎叶图
x=rnorm(380);y=x+rnorm(380)                    #构造线性关系
plot(y~x)
fit=lm(y~x)
abline(fit,col="blue")              #加入拟合线
demo(graphics)                  #好多图片哦

#实践 7  复数运算与求函数极值
(2+4i)^-3.5+(2i+4.5)*(-1.7-2.3i)/((2.6-7i)*(-4+5.1i))      #复数运算
#下面构造一个 10 维复向量，实部和虚部均为 10 个标准正态样本点：
```

```
(z<-complex(real=rnorm(10),imaginary=rnorm(10)))
complex(re=rnorm(3),im=rnorm(3))                    #3 维复向量
Re(z)          #实部
Im(z)          #虚部
Mod(z)         #模
Arg(z)         #辐角
choose(8,2)    #组合 c82=28
factorial(6)   #6 的阶乘 6！=720
#解方程
f=function(x)x^3-2*x-1
uniroot(f,c(0,2))          #迭代求根
#如果知道根为极值
f=function(x)x^2+2*x+1     #定义一个二次函数
optimize(f,c(-2,2))        #在区间-2，2 之间求极值

#实践 8 字符型向量
a=factor(letters[1:10])    #letters:小写字母的向量，LETTERS:大写字母
a[3]="w"                   #不行！会出警告
a=as.character(a)          #转换一下
a[3]="w"                   #可以了
a;factor(a)                #两种不同类型的向量

#实践 9 数据输入输出
x=scan()                              #屏幕输入，可键入或粘贴，多行输入在空行后按 enter
#1.5 2.6 3.7,2.1 8.9 12,-1.2 -4
x=c(1.5, 2.6, 3.7, 2.1, 8.9, 12, -1.2, -4)      #等价于上面
w=read.table(file.choose(),header=T)
apropos("wd")
setwd("bootcamp:/2017 多元统计分析")   #建立工作路径
(x=rnorm(20))                         #给 x 赋值 20 个标准正态数据
write(x,"test.txt")                   #把数据写入文件
y=scan("test.txt");y                  #扫描文件数值数据到 y
y=iris;y[1:5,]                        #iris 自带数据
str(y)                                #显示数据的内部结构
write.table(y,"test.txt",row.names=F)  #把数据写入文本文件
w=read.table("test.txt",header=T)      #读带有变量名的数据
str(w)
write.csv(y,"test.csv")                #把数据写入 csv 文件
read.csv("test.csv")                   #读入 csv 数据文件
str(v)                                 #汇总
data=read.table("clipboard")           #读入剪贴板的数据

#实践 10 序列
(z=seq(-1,10,length=100))      #从-1 到 10 等间隔的 100 个数组成的序列
```

```
z=seq(-1,10,len=100)          #从-1 到 10 等间隔的 100 个数组成的序列
(z=seq(10,-1,-0.1))           #从 10 到-1，间隔为-0.1 的序列
(x=rep(1:3,3))                #重复 123 三次：123123123
(x=rep(3:5,1:3))              #重复 345 依次 123 次：344555
(x=rep(c(1,10),c(4,5)))       #重复 1，10 分别 4 次 5 次：1 1 1 1 10 10 10 10 10
(w=c(1,3,x,z))                #合并向量
w[3]                          #向量 w 中第三个数值
x=rep(0,10);z=1:3;x+z         #向量的加法，如果长度不同，会给出警告
x*z
rev(x)
z=c("no cat","has ","nine","tails")      #字符向量
z[1]=="no cat"                #双等号为逻辑等式
z=1:5
z[7]=8;z                      #缺失第六个数
z=NULL                        #清空 z 值
z[c(1,3,5)]=1:3;z
rnorm(10)[c(2,5)]
z[-c(1,3)]                    #去掉第 1、3 个元素
z=sample(1:100,10);z
which(z==max(z))              #给出最大值下标

#实践 11  矩阵
x=sample(1:100,12);x
all(x>0);all(x!=0);any(x>0);(1:10)[x>0]     #逻辑符号的应用
diff(x)                #差分
diff(x,lag=2)          #2 步差分
diff(x,lag=3,2)        #3 步两阶差分
x=matrix(1:20,4,5,byrow=T)          #矩阵的构造，按行排列
t(x)
x=matrix(sample(1:100,20),4,5)
2*x;x+5
y=matrix(sample(1:100,20),5,4)
x+t(y)
(z=x%*%y)              #矩阵乘法
z1=solve(z)            #解方程
z1%*%z
round(z1%*%z,5)        #四舍五入
b=solve(z,1:4);b

#实践 12  矩阵
x=matrix(rnorm(24),4,6)
nrow(x);ncol(x);dim(x)
x[c(2,1),];x[,c(1,3)];x[2,1];x[x[,1]>0,1]
sum(x[,1]>0);sum(x[,1]<=0)
```

```
x[,-c(1,3)]        #去掉第 1、3 列的 x
diag(x)            #X 矩阵对角元
diag(1:5)          #对角线为 12345 的对角矩阵
diag(5)            #五维单位阵
x[-2,-c(1,3)]      #去掉第 2 行及第 1、3 列的 x
x[x[,1]>0&x[,3]<=1,1]     #第一列大于 0 且第三列小于等于 1 的行对应的第一列元素
x[x[,2]>0|x[,1]<.51,1]    #第 2 列大于 0 或者第 1 列小于 0.51 的行对应的第一列元素
x[!x[,2]<.51,1]           #第 1 列中相应于第 2 列大于等于 0.51 的元素
apply(x,1,mean)           #对行（第一维）求均值
apply(x,2,sum)            #对列（第二维）求均值
x=matrix(rnorm(24),4,6)
x[lower.tri(x)]=0;x       #得到上三角矩阵,#下三角 x[upper.tri(x)]=0

#实践 13  高维数组
x=array(runif(24),c(4,3,2));x     #用 24 个服从均匀分布的样本点构造 4*3*2 的三维数组
is.matrix(x)
dim(x)
is.matrix(x[1,,])
x=array(1:24,c(4,3,2));x
x[c(1,3),,]
x=array(1:24,c(4,3,2))
apply(x,1,mean);apply(x,1:2,sum)
apply(x,c(1,3),prod)              #对部分维做求乘积运算

#实践 14  矩阵与向量之间的运算
x=matrix(1:20,5,4);x
sweep(x,1,1:5,"*")        #把向量 1：5 的每个元素乘到每一行
sweep(x,2,1:4,"+")        #把向量 1：4 的每个元素加到每一列
x*1:5
#下面把 x 标准化，即每一个元素减去该列均值，除以该列标准差
(x=matrix(sample(1:100,24),6,4));(x1=scale(x))
#对 x 矩阵进行标准化：每列减去均值然后除以标准差
(x-mean(x))/sd(x)
#这个标准化和 scale 不同，scale 按列进行，后面减 de 是所有元素的平均值除的是所有元素标准差
(x2=scale(x,scale=F))#进行中心化：即每一个元素减去该列均值
(x3=scale(x,center=F))
round(apply(x1,2,mean),14)
apply(x1,2,sd)
round(apply(x2,2,mean),14);apply(x2,2,sd) #
round(apply(x3,2,mean),14);apply(x3,2,sd) #

#实践 15  缺失值处理与数据的合并
airquality                #有缺失值（NA）的 R 自带数据
```

```
names(airquality)
complete.cases(airquality)                    #判断每行有没有缺失值
which(complete.cases(airquality)==F)    #有缺失值得行号
sum(complete.cases(airquality))           #完整观测值的个数
na.omit(airquality)                           #删除缺失值数据
#附加，横或竖合并数据：spend,cbind,rbind
x=1:10;x[12]=3
(x1=append(x,77,after=5))
cbind(1:5,rnorm(5))
rbind(1:5,rnorm(5))
cbind(1:3,4:6)
rbind(1:3,4:6)
(x=rbind(1:5,runif(5),runif(5),1:5,7:11))
x[!duplicated(x),]
unique(x)                                     #去掉矩阵重复的行

#实践 16 list
#list 可以是任何对象（包括 list 本身）的集合
z=list(1:3,Tom=c(1:2,a=list("R",letters[1:5]),w="hi!"));z
z[[1]];z[[2]]
z$T
z$T$a2
z$T[[3]]
z$T$w

#实践 17  条形图和表
x=c(1,2,3,3,2,2,4,4,4,5,6,7,2,3,4,6,9,10,11,2,1,5,6)
barplot(x,col = "lightblue")
table(x)
barplot(table(x),col = "lightblue")
barplot(table(x)/length(x),col = "lightblue")
table(x)/length(x)

#实践 18  形成表格
library(MASS)   #载入软件包 MASS
quine    #MASS 中所带数据
attach(quine) #把数据变量的名字放入内存
#下列语句产生从该数据得到的各种表格
table(Age)
table(Sex,Age)
tab=xtabs(~Sex+Age,quine)
unclass(tab)
tapply(Days,Age,mean)
detach(quine) #attach 的逆运行
```

```
#实践 19  如何写函数
myfun=function(n){
  z=2 #从最小的素数 2 开始
  for (i in 2:n) {
    if(any(i%%2:(i-1)==0)==F){
      #用 i 整除 2 到 i-1 的所有数,如果所有这些余数都不等于 0,那么 i 就是素数，将它
加进 z 向量中
      z=c(z,i)
    }
  }
  return(z)
}
myfun(100)

#实践 20  画图
x=seq(-3,3,len=20);x   #产生数据
y=dnorm(x) ;y       #x 为样本点的正态分布概率值作为 y
w=data.frame(x,y) #合并 x，y 成为数据 w
par(mfcol=c(2,2))   #准备画四个图的地方：2 行 2 列
plot(y~x,w,type="l",main="正态密度函数")
plot(y~x,w,type="o",main="正态密度函数")
plot(y~x,w,type="b",main="正态密度函数")
par(mfcol=c(1,1)) #取消 par(mfcol=c(2,2))

#实践 21  色彩与符号等的调节
plot(1,1,xlim=c(1,7.5),ylim=c(0,5),type="n")#画出框架
#在 plot 命令后面追加点（追加线可用 lines 函数）
points(1:7,rep(4.5,7),cex=seq(1,4,l=7),col=1:7,pch=0:6)
text(1:7,rep(3.5,7),labels=paste(0:6,letters[1:7])
     ,cex=seq(1,4,l=7),col=1:7)   #在指定位置加文字
points(1:7,rep(2,7),pch=(0:6)+7)   #点出符号 7 和 13
text((1:7)+0.25,rep(2,7),paste((0:6)+7)) #加符号号码
points(1:7,rep(1,7),pch=(0:6)+14)   #点出符号 14 和 20
text((1:7)+0.25,rep(1,7),paste((0:6)+14)) #加符号号码
#关于符号形状，大小，颜色以及其他画图选项的说明可以用"?par"来查看
```

实验题目 2 聚类分析

实验目的

了解、掌握，并能熟练地对数据进行相关的变换处理，掌握中心化变换、标准化变换，极差正规化变换、对数变换等，对系统聚类法与动态聚类法进行实践运用，并作相关分析，提高对数据聚类的认识。

实验工具

计算机，Windows，R 软件

实验内容

1. 对四省市消费支出数据进行至少四种数据变化；

City	肉质及制品	住房	医疗保健	交通与通讯	文娱用品及服务
北京	563.51	227.78	147.76	235.99	510.78
上海	678.92	365.07	112.82	301.46	465.88
陕西	237.38	174.48	119.78	141.07	245.57
甘肃	253.41	156.13	102.96	108.13	212.2

2. 聚类分析（二选一）。

1) 全国区域经济的系统聚类和动态聚类分析[王斌会]p150；

2) 考虑海南省 19 只 A 股 2012 年 6 月 29 日的数据，选择 11 个指标如表[杨虎]p7，为了减少数量级的影响，先对数据作标准化处理，然后作系统聚类和动态聚类，类的个数选择为 5[杨虎]p28

实验指导

1. 对四省市消费支出数据进行至少四种数据变化

```
#中心化变换:原始数据减去对应变量的均值
mean=t(as.matrix(apply(data,2,mean))) #将 4 个均值转换为行向量
mean=matrix(rep(mean,4),ncol = 5,byrow = T) #将均值的行向量重复 4 行
data1=data-mean;data1
scale(data,center = T,scale = F)
#标准化变换：中心化后，除以各变量的标准差
sd=t(as.matrix(apply(data, 2, sd)))
sd=matrix(rep(sd,4),ncol = 5,byrow = T)
data2=data1/sd;data2
scale(data)
#极差正规化变换：原始数据减去该变量的最小值，再除以极差
d=apply(data, 2, max)-apply(data,2,min)
d=t(as.matrix(d))
d=matrix(rep(d,4),ncol = 5,byrow = T);d #极差
min=t(as.matrix(apply(data,2,min)))
min=matrix(rep(min,4),ncol = 5,byrow = T);min #最小值
```

```
data3=(data-min)/d;data3
#对数变换：对原始数据取以 e 为底的对数
data4=log(data);data4
#小数定标规范化:将极差正规化变换后的数据进行小数定标
options(digits = 4)
data5=data3;data5
```

2、全国区域经济的系统聚类和动态聚类分析[王斌会]p150

```
library(openxlsx)
data=read.xlsx('D:\\多元统计分析\\作业\\mvcase5.xlsx',sheet=7,rowNames = T)
#标准化数据
sc_data=scale(data)
#系统聚类
d=dist(sc_data)        #距离矩阵,默认欧氏距离
#使用 ward.D2 法计算类与类之间的距离
hc=hclust(d,method = "ward.D2")
#画出聚类图;给定分类数为 4,加入分类框;返回各地区所属的类别
plot(hc);rect.hclust(hc,4)
cutree(hc,4)

#动态聚类——k-means 聚类
kmeans(sc_data,4)$cluster
```

实验题目 3 判别分析

实验目的

对相关的数据进行判别分析，掌握不同的具体判别方法的实践与应用，加深对数据判别方法的理解。

实验工具

计算机，Windows，R 软件

实验内容

企业财务状况的判别分析（王斌会 p126）

实验指导

企业财务状况的判别分析

```
library(readxl)
library(MASS)
data=read_xlsx('D:\\多元统计分析\\作业\\mvcase5.xlsx',sheet=6)
#做 Fisher 线性判别
#构建线性判别模型
l=lda(G~CF_TD+NI_TA+CA_CL+CA_NS,data=data);l
pl=predict(l)              #根据线性判别模型预测样本所属类别
G1=pl$class                #预测的所属类别结果
d1=data.frame(data$G,G1);d1      #将真实的类别与预测的类别组成数据框
tab1=table(data$G,G1);tab1       #判别矩阵
sum(diag(prop.table(tab1)))      #该线性模型的判对率

#做 Fisher 非线性判别
#构建非线性判别模型
q=qda(G~CF_TD+NI_TA+CA_CL+CA_NS,data=data);q
pq=predict(q)                  #根据非线性判别模型预测样本所属类别
G2=pq$class                    #非线性模型预测出的所属类别结果
d2=data.frame(data$G,G2)
tab2=table(data$G,G2);tab2         #判别矩阵
sum(diag(prop.table(tab2)))        #该非线性模型的判对率
```

实验题目 4 主成分分析

实验目的

了解主成分分析的统计思想和实际意义，以及它的数学模型和在二维空间上的几何解释；能熟练地利用 R 语言对数据进行相关分析，提取出主成分；能够自己编程解决实际问题并给出分析报告。

实验工具

计算机，Windows，R 软件

实验内容

1. 主成分分析——洛杉矶街区数据

数据网站：

http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_LA_Neighborhoods_Data ，一共有 110 个街区，15 个变量。

2. 主成分回归[于秀林 p199]

实验指导

1.洛杉矶街区数据——分步编程

```
library(openxlsx)
data=read.xlsx("D:\\多元统计分析\\作业\\第五章 主成分分析 数据.xlsx",sheet=2,rowName=TRUE)
data$人口密度=data$人口/data$面积
data=data[,-c(11,12,13,14)]
#计算样本相关系数矩阵
sigma=cor(data)
#计算相关系数矩阵的特征值和特征向量
lambda=eigen(sigma)
lambda$values          #相关系数矩阵的特征值
#根据特征值计算方差贡献率和累计贡献率
options(digits = 4)
contribute_ratio=data.frame(comp=c(1:11),贡献率=lambda$values/sum(lambda$values))
contribute_ratio
contribute_ratio$累计贡献率=cumsum(contribute_ratio$贡献率)
contribute_ratio
#以特征值为纵坐标画碎石图
plot(x=c(1:11),lambda$values,'o',lty=1,col='chartreuse1',xlab='主成分')
abline(h=1,lty=3)
text(x=c(1:11),lambda$values,cex=0.7,round(lambda$values,2),pos = 1)
#主成分载荷矩阵
lambda$vectors[,1:4]                    #排名前 4 的特征值对应的特征向量
loading=data.frame(lambda$vectors)   #生成载荷矩阵
```

```
#改一下变量名
colnames(loading)=c("comp.1","comp.2","comp.3","comp.4","comp.5","comp.6","comp.7","comp.8","comp.9","comp.10","comp.11")
loading[,1:4]
#主成分得分
#由于数据框中的行数据计算时有可能出错,将原始数据转换为矩阵并存在另一个变量中,方便后续计算
data1=as.matrix(data)
#将原始数据框复制给 scores，主要是利用原始数据框的框架，后续会把其中数据删除，只留下主成分得分
scores=data
for(i in 1:5){
  scores[,i]=scale(data1)%*%loading[,i]
}
scores=scores[,-c(6:11)] #删除其他数据，只留下得分
colnames(scores)=c("comp.1","comp.2","comp.3","comp.4","comp.5")
head(scores)

#画主成分载荷图,前 4 个主成分,每两个作为一组坐标轴
par(mfrow=c(1,2))
plot(loading[,1:2],type="n",main="loadings")
text(loading[,1],loading[,2],cex=0.9,c("收入","API","种族","年龄","有房家庭比例","复员军人比例","亚裔比例","非裔比例","拉美裔比例","欧裔比例","人口密度"))
abline(h=0,v=0,lty=3)
plot(loading[,3:4],type="n",main="loadings")
text(loading[,3],loading[,4],cex=0.9,c("收入","API","种族","年龄","有房家庭比例","复员军人比例","亚裔比例","非裔比例","拉美裔比例","欧裔比例","人口密度"))
abline(h=0,v=0,lty=3)

#主成分得分排序
final_score=0
for (i in 1:5) {
  final_score=final_score+contribute_ratio$贡献率[i]*scores[,i]
}
colnames(final_score)='score'
head(final_score[order(-final_score),])
```

2.洛杉矶街区数据——princomp 函数

```
ca=princomp(scale(data),cor = T)
summary(ca)
options(digits = 4)
ca$sdev
ca$loadings[,1:3]
lambda$values^0.5
loading[,1:3]
```

3.主成分回归[于秀林 p199]

```
data=data=read.xlsx("D:\\多元统计分析\\作业\\第五章　主成分分析　数据.xlsx",sheet=3,rowName=TRUE)
#直接线性回归
fit1=lm(Y~x1+x2+x3,data=data)
summary(fit1)

#主成分回归
ca=princomp(data[,1:3],cor=T) #线性回归时未进行标准化，因此在此选择协方差阵
summary(ca) #选择前两个主成分
Comp.1=ca$scores[,1];Comp.2=ca$scores[,2]
fit2=lm(data$Y~Comp.1+Comp.2)
summary(fit2)
#做变换，得到原坐标下的表达式
beta<-coef(fit2); A<-loadings(ca)
x.bar<-ca$center; x.sd<-ca$scale
coef<-(beta[2]*A[,1]+ beta[3]*A[,2])/x.sd
beta0 <- beta[1]- sum(x.bar * coef)
c(beta0, coef)
```

实验题目 5 因子分析

实验目的

了解因子分析的目的和实际意义，熟悉因子分析数学模型建模的假设条件和各个分量的实际统计意义；掌握由主因子方法估计因子载荷的上机操作步骤，利用 R 语言编程解决实际问题中的因子分析问题，同时给出初步的统计分析报告。

实验工具

计算机，Windows，R 软件

实验内容

对 31 个省市自治区 8 个家庭平均每人全年消费性支出数据的人均消费水平作分析评价，并根据因子得分和综合得分对各省市自治区的人均消费水平作因子分析[王斌会]p189,数据见文件 mvstats4 王斌会--例子数据 d7.2。

实验指导

```
library(openxlsx)
data=read.xlsx("D:\\多元统计分析\\作业\\第五章 主成分分析 数据.xlsx",sheet=1,rowNames=TRUE)
#确认数据是否适合做因子分析
cor(data)
msa.bartlett(data)                    #p 值很小，说明数据适宜做因子分析
#基于主成分估计的因子函数
fa=msa.fa(data,2,rotation = 'none')
#因子旋转
fa1=msa.fa(data,2,rotation = 'varimax')
#旋转前后的载荷矩阵对比
fa$loadings;fa1$loadings
#旋转前后的因子得分对比
fa$scores;fa1$scores
#因子得分信息图
plot(fa1$scores);abline(h=0,v=0,lty=3)
text(fa1$scores,labels = rownames(data))
#旋转前后综合得分及排名
fa$ranks;fa1$ranks
```

实验题目 6 对应分析

实验目的

运用对应分析生成的图形，对相关的数据规律或市场规律做进一步的分析与预测。在每个变量划分多个类别的情况下揭示出变量间的内在联系，掌握对应分析的方法与实战技巧，并能熟练将其方法运用于市场细分、产品定位、品牌形象及满意度研究的方向上。

实验工具

计算机，Windows，R 软件

实验内容

根据网站评价 Web 站点 Alexa 所提供的评价指标数据，选取 5 个指标作为媒体网站评价标准：流量、访问量、被连接数、速度、浏览页面数。对这些媒体网站进行评价，分析媒体网站的定位。

实验指导

```
library(openxlsx)
data=read.xlsx("D:\\多元统计分析\\作业\\mvcase5.xlsx",sheet=10,rowNames=T)
#进行对应分析
library(ca)
result=ca(data)
summary(result)
result$rowcoord #行坐标
result$colcoord #列坐标
plot(result)
```

实验题目 7 典型分析

实验目的

了解典型相关分析的目的和统计思想，以及典型相关的实际意义。了解计算机软件程序中有关典型相关分析的基本内容，能运用 R 语言进行典型相关分析。

实验工具

计算机，Windows，R 软件

实验内容

某健康俱乐部对 20 名中年人测量了三个生理指标：体制 X1、腰围 X2 和脉搏 X3；同时也测量了三个训练指标：引体向上次数 y1、仰卧起坐次数 y2 和跳跃次数 y3，试作生理指标和训练指标的典型相关分析。

实验指导

1.分步编程

```
library(openxlsx)
data=read.xlsx("D:\\多元统计分析\\实验\\第八章  典型相关分析数据.xlsx")
data=scale(data)      #数据标准化
#进行矩阵的分块
R=cov(data)
R11=R[1:3,1:3]
R12=R[1:3,4:6]
R21=R[4:6,1:3]
R22=R[4:6,4:6]
A=solve(R11)%*%R12%*%solve(R22)%*%R21
B=solve(R22)%*%R21%*%solve(R11)%*%R12
#求 A B 的特征根
values_A=eigen(A)$value;values_A
values_B=eigen(B)$value;values_B
#典型相关系数为
(coef=sqrt(values_A))
#A B 特征值对应的特征向量,即典型载荷
(vectors_A=eigen(A)$vectors)
(vectors_B=eigen(B)$vectors)

#---------典型相关系数的检验---------
#第一个典型相关系数的检验
chs=qchisq(0.95,df=(3-1+1)*(3-1+1));chs        #0.05 的显著性水平下的临界值
n=nrow(data)
delta_0=cumprod(1-values_A)[3]
Q0=-(n-1-0.5*(3+3+1))*log(delta_0);Q0
```

```
#Q0 小于临界值，不能拒绝原假设，认为第 1 个典型相关系数不显著
#第二个典型相关系数的检验
chs=qchisq(0.95,df=(3-2+1)*(3-2+1));chs        #0.05 的显著性水平下的临界值
n=nrow(data)
delta_1=cumprod((1-values_A)[2:3])[2]
Q1=-(n-2-0.5*(3+3+1))*log(delta_1);Q1
#Q1 小于临界值，不能拒绝原假设，认为第 2 个典型相关系数不显著
#第三个典型相关系数的检验
chs=qchisq(0.95,df=(3-3+1)*(3-3+1));chs #0.05 的显著性水平下的临界值
n=nrow(data)
delta_2=(1-values_A)[3]
Q2=-(n-3-0.5*(3+3+1))*log(delta_1);Q2
#Q2 小于临界值，不能拒绝原假设，认为第 3 个典型相关系数不显著
#说明 3 个典型相关系数均不显著，所以就不需要做进一步的典型相关分析了
```

2. 使用 cancor 函数

```
result=cancor(data[,1:3],data[,4:6])
result
#虽然两种方法下的典型载荷不同，但是它们的比值是常数
vectors_A/result$xcoef
vectors_B/result$ycoef
```

实验题目 8 广义线性模型

实验目的

针对因变量和自变量的取值性质，了解统计模型的类型。掌握数据的分类与模型的选择方式，并对广义现性模型和一般线性模型有初步的了解，掌握对应的上机实现。

实验工具

计算机，Windows，R 软件

实验内容

关于 40 个不同年龄 age 和性别 sex 的人对某项服务产品的观点 y 的数据进行 Logistic 回归。

实验指导

```
library(openxlsx)
data=read.xlsx("D:\\多元统计分析\\作业\\mvcase5.xlsx",sheet=5)
result=glm(y~sex+age,family = binomial,data = data)
summary(result)
pd=predict(result,data)          #根据已有 sex,age 数据进行预测
p=exp(pd)/(1+exp(pd));p      #持观点 1 的概率
data$p=p
attach(data)
plot(sex,p) #所持观点与性别的关系
plot(age,p) #所持观点与年龄的关系
```