华南理工大学

# 实 验 报 告

| | |
|---|---|
| 课程名称： | R 统计软件 |
| 学生姓名： | 黄庆昌 |
| 学生学号： | 201630450061 |
| 学生专业： | 数学与应用数学（统计学方向） |
| 开课学期： | 2018-2019 年第 2 学期 |

**数学学院**

**2019 年 5 月 制**

# 目 录

# 实验一 R 语言基础

| | | | |
|---|---|---|---|
| **地　　点：** | 4　楼　4105 房； | **实验台号：** | 个人电脑 |
| **实验日期与时间：** | 2019 年 4 月 26 日 | **评　　分：** | |
| **预习检查纪录：** | | **实验教师：** | 龙卫江 |
| **电子文档存放位置：** | C:\ex | | |
| **电子文档文件名：** | 201630450061 黄庆昌 实验一.doc | | |

**批改意见**

# 实验一、R 语言基础

## 一、实验目的

1 理解 R 语言的主要数据类型与数据结构，基本运算和函数；
2 掌握数据管理基本知识与操作；
3 实现一些功能的 R 编程。要求学生理解理解 R 语言的主要数据结构、数据管理基本技术、R 函数的使用和编码。

## 二、实验环境

R-3.5.3，R 软件包 xlsx，Excel 表格文件 College.xls

## 三、实验原理

1、数据框：数据框可通过函数 data.frame(col1,col2,col3,…)创建，其中的列向量 col1、col2、col3 等可为任何类型，（如字符型、数值型或逻辑型），每一列的名称由函数 names 指定。数据框是一种特殊的列表对象，有一个值为“data.frame”的 class 属性，各列表成员必须是向量、因子、数值型矩阵、列表或其他数据框.向量、因子成员为数据框提供一个变量，非数值型向量会被强制转换为因子，而矩阵、列表、数据框这样的成员为新数据框提供了和其列数、成员数、变量数相同个数的变量、作为数据框变量的向量、因子或矩阵必须具有相同的长度。
2、外部数据的读入：xlsx 包可以用来对 Excel 97/2000/XP/2003/2007 文件进行读取、写入和格式转换。函数 read.xlsx()导入一个工作表到一个数据框中。最简单的格式是 read.xlsx(file，n)，其中 file 是 Excel 工作簿的所在路径，n 则为要导入的工作表序号。
3、数据库连接的方法：（1）在 R 中能通过 RODBC 包访问数据库，这种方式允许 R 连接到任意一种拥有 ODBC 驱动的数据库；（2）DBI 包为访问数据库提供了一个通用且一致的客户端接口，构建于这个框架之上的 RJDBC 包提供了通过 JDBC 驱动访问数据库的方案。
4、数据框元素的引用：（1）可以使用下标或下标向量，也可以使用列名或由列名构成的向量；（2）数据框的各变量也可以按列表引用（即用双括号[[]]或$符号引用）
5、常见的函数及功能：
mean():求平均数；median():求中位数；sd():求标准差；var():求方差
rnorm():生成随机数；sample(x,size):随机抽样

## 四、实验内容

（一）.数据处理与数据管理。

利用 College.xls 数据集。

1 数据读入：在 C 盘建目录 ex,将 College.xls 拷入 C:\ex 下。

1.A) 将 College.xls 数据读入 R 中对象 College；需要什么包？查 College 的结构。

答：需要 xlsx 包。College 为一数据框，结构如图 1 所示：

```
> library(xlsx)
> a<-read.xlsx("C:/ex/College.xls",1)
> View(a)
> str(a)
'data.frame':   777 obs. of  19 variables:
 $ NA.        : Factor w/ 777 levels "Abilene Christian University",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Private    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ Apps       : num  1660 2186 1428 417 193 ...
 $ Accept     : num  1232 1924 1097 349 146 ...
 $ Enroll     : num  721 512 336 137 55 158 103 489 227 172 ...
 $ Top10perc  : num  23 16 22 60 16 38 17 37 30 21 ...
 $ Top25perc  : num  52 29 50 89 44 62 45 68 63 44 ...
 $ F.Undergrad: num  2885 2683 1036 510 249 ...
 $ P.Undergrad: num  537 1227 99 63 869 ...
 $ Outstate   : num  7440 12280 11250 12960 7560 ...
 $ Room.Board : num  3300 6450 3750 5450 4120 ...
 $ Books      : num  450 750 400 450 800 500 500 450 300 660 ...
 $ Personal   : num  2200 1500 1165 875 1500 ...
 $ PhD        : num  70 29 53 92 76 67 90 89 79 40 ...
 $ Terminal   : num  78 30 66 97 72 73 93 100 84 41 ...
 $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
 $ perc.alumni: num  12 16 30 37 2 11 26 37 23 15 ...
 $ Expend     : num  7041 10527 8735 19016 10922 ...
 $ Grad.Rate  : num  60 56 54 59 15 55 63 73 80 52 ...
>
```

**图 1：College 结构**

1.B) 将 College.xls 存为将 College.csv，将 College.csv 读入 R 中对象 College1。

答：代码如下所示：

```
library(xlsx)
a<-read.xlsx("C:/ex/College.xls",1)
str(a)
write.csv(a,"C:/ex/College.csv",row.names = FALSE)
College1<-read.csv("C:/ex/College.csv")
```

2.数据操作：

2.A)在 R 中将 College 中学校为私立且毕业率>百分之六十的记录存入对象 CollegeEx,并为 CollegeEx 添加一个命名为报录比率的变量，命名为 AAR。

2.B)将添加 AAR 后的 CollegeEx 抽取 Apps,Accept,Books,AAR 形成的子集存为 CollegeEx0.

2.C)提取 CollegeEx0 中前 5 笔的 Apps,AAR 数据，并存入 CollegeEX0.

2.D)将 CollegeEx0 存为 C:\ex 下 csv 格式的外部文件 CollegeEx0.csv.

答：代码如下所示：

（其中将 ARR 变量赋值为全 1 变量以便插入数据）

```
CollegeEx<-College1[College1$Private=="Yes"&College1$Grad.Rate>60,]
CollegeEx$AAR<-1
CollegeEx0<-CollegeEx[c("Apps","Accept","Books","AAR")]
CollegeEX0<-CollegeEx0[c("Apps","AAR")][0:5,]
write.csv(CollegeEx0,"C:/ex/CollegeEx0.csv",row.names = FALSE)
```

3.基本函数使用：

3.A)对 CollegeEx0，给出 Books 的均值、方差、标准差，求 Books 的离差平方和；给出 CollegeEx0 描述统计概要，针对 Books 相应地解释概要数据的统计意义。

答：Books 的均值为 541.4623，方差为 25112.5，标准差为 158.4692，离差平方和为 25049.4，CollegeEx0 的统计概要如图 2 所示：

```
> College1<-read.csv("C:/ex/College.csv")
> CollegeEx<-College1[College1$Private=="Yes"&College1$Grad.Rate>60,]
> CollegeEx$AAR<-1
> CollegeEx0<-CollegeEx[c("Apps","Accept","Books","AAR")]
> CollegeEX0<-CollegeEx0[c("Apps","AAR")][0:5,]
> b<-CollegeEx0$Books
> mean(b)
[1] 541.4623
> var(b)
[1] 25112.5
> sd(b)
[1] 158.4692
> sumofsquare<-var(b)*((length(b)-1)/length(b))
> sumofsquare
[1] 25049.4
> summary(CollegeEx0)
      Apps             Accept          Books             AAR
 Min.   :  141.0   Min.   :  118   Min.   : 250.0   Min.   :1
 1st Qu.:  860.2   1st Qu.:  665   1st Qu.: 450.0   1st Qu.:1
 Median : 1467.5   Median : 1083   Median : 500.0   Median :1
 Mean   : 2371.4   Mean   : 1518   Mean   : 541.5   Mean   :1
 3rd Qu.: 2792.8   3rd Qu.: 1830   3rd Qu.: 600.0   3rd Qu.:1
 Max.   :20192.0   Max.   :13007   Max.   :2000.0   Max.   :1
> |
```

**图 2：CollegeEx0 描述统计概要**

对 CollegeEx0 中的 Books 列，可知其最小值为 250.0，最大值为 2000.0，平均值为 541.5，众数为 500.0，第一分位数为 450.0，第三分位数为 600.0。

（二）.数据操作与运算。

1.准备数据

设置种子数 40104，生成容量为 40 的 N(1,3)的样本值 x;

设置种子数 40105，生成容量为 20 的 1:100 等可能取值的样本值 y;

将 rep(1:2,2),rep(1:3,c(2,1,1)),c(3,2)组成向量 z.

答：x、y、z 的生成如图 3 所示：

```
> set.seed(40104)
> x<-rnorm(40,1,sqrt(3))
> set.seed(40105)
> y<-sample(1:100,size = 20)
> z<-c(rep(1:2,2),rep(1:3,c(2,1,1)),c(3,2))
> x
 [1]  2.7540422  7.4526937  2.6438413 -0.6595340 -0.9348761  1.4250726 -0.3189949  4.2158913  0.3506054  2.1778696 -0.1604569
[12] -0.4343168  1.5562401 -2.3091386  1.9983579  1.6302481 -0.7380569  1.6744792 -2.0698902  0.5558830 -1.2619525  1.3916675
[23]  0.7161787  2.8001523  1.9461710  3.6098024  0.7230238  2.2285523 -1.2347489  2.4837916  2.2510690  1.1826758 -0.8105294
[34]  1.2531797 -2.4497619  1.0640266  1.7276917  0.1680047  3.4297343  2.5691464
> y
 [1] 81 91 26 24 31 96 18 20  9 12 36 69 56 97 61 44 11 55 87 88
> z
 [1] 1 2 1 2 1 1 2 3 3 2
>
```

**图 3：x、y、z 的生成**

所用代码：

```
set.seed(40104)
x<-rnorm(40,1,sqrt(3))
set.seed(40105)
y<-sample(1:100,size = 20)
z<-c(rep(1:2,2),rep(1:3,c(2,1,1)),c(3,2))
```

2. 转换

   由 x 按行顺序填充生成一个 10*4 的矩阵 xx;

   由 y 按列顺序填充生成一个 10*2 的数据框 yy;

   由 z 按列顺序填充生成一个 10*1 的数据框 zz;

   按列合并 xx,yy,zz 为数据框 W.

   为 W 的列命名：x,y,z,u,v,w,t

3. 将 W 的第 1、2、3、8 行，第 1、3、6 列的子块组成 V.

   给出 W 的第 2 列的作为向量的模长；

   给出 V 的行和向量、列和向量。

理解所用到的函数和操作，理解功用

答：2 和 3 的代码如下所示：

```
xx<-matrix(x,nrow = 10,byrow = TRUE)
yy<-data.frame(array(y,c(10,2)))
zz<-data.frame(z)
W<-cbind(cbind(data.frame(xx),yy),zz)
colnames(W)<-c("x","y","z","u","v","w","t")
norm(as.matrix(W[,2]),type="2")
rowSums(V)
colSums(V)
```

W 的第 2 列的作为向量的模长为 9.653389，V 的行和向量为 rowSums(V)，V 的列和向量为 colSums(V)。结果如图 4 所示：

```
> #二.数据操作与运算
> set.seed(40104)
> x<-rnorm(40,1,sqrt(3))
> set.seed(40105)
> y<-sample(1:100,size = 20)
> z<-c(rep(1:2,2),rep(1:3,c(2,1,1)),c(3,2))
> xx<-matrix(x,nrow = 10,byrow = TRUE)
> yy<-data.frame(array(y,c(10,2)))
> zz<-data.frame(z)
> W<-cbind(cbind(data.frame(xx),yy),zz)
> #按列合并xx,yy,zz为数据框W
> colnames(W)<-c("x","y","z","u","v","w","t")
> V<-W[c(1,2,3,8),c(1,3,6)]
> norm(as.matrix(W[,2]),type="2")
[1] 9.653389
> rowSums(V)
       1        2        3        8
41.39788 67.74613 56.19015 56.01632
> colSums(V)
          x           z           w
  0.9350226   4.4154586 216.0000000
>
```

图 4：转换

## 五、实验总结

通过这次实验，使我对 R 中基本数据结构、数据库的连接方法、对外部文件的读写操作、数据框的创建及数据框元素的访问、R 中常用统计函数的使用等有了更深的了解，在实际操作中锻炼了打代码的能力。

# 实验二 R 数据预处理、描述统计及绘图的实现

| | | | |
|---|---|---|---|
| **地　　点：** | 4　楼　4105 房； | **实验台号：** | 个人电脑 |
| **实验日期与时间：** | 2019 年 5 月 3 日 | **评　　分：** | |
| **预习检查纪录：** | | **实验教师：** | 龙卫江 |
| **电子文档存放位置：** | C:\ex | | |
| **电子文档文件名：** | 201630450061 黄庆昌 实验二.doc | | |

**批改意见**

# 实验二、数据预处理、描述统计及绘图的实现

## 一、实验目的

1 熟悉理解数据预处理技术，掌握其 R 实现；

2 掌握描述统计的 R 实现；

3 熟悉理解 R 绘图的原理和主要函数。

要求学生理解数据预处理技术、描述统计、数据可视化展示的 R 实现，掌握相关 R 函数的使用和熟悉基本功能的编码实现。

## 二、实验环境

R 软件，必要的 R 软件包，鸢尾花等数据集(提供)

## 三、实验原理

数据预处理、描述性分析和绘图是我们开展数据分析工作的重要一步，所涉及的描述统计知识是我们理解数据、展示数据的基础内容，其基本原理可参考前修的数理统计课程的基础内容。R 软件在数据整形、缺失处理、汇总，以及在描述统计和图形展示方面，都提供了许多函数和库，掌握这些内容，有助于我们对实际问题选用恰当工具、通过编程实现特定功能需求，并能恰当地解释数据所蕴含的信息。

## 四、实验内容

1、简要叙述数据整形常用的库及函数的主要功能、常用的汇总函数的主要功能、描述统计的常用方法、R 中常用的绘图函数及功能

答 1：

（1）数据整形可以使用 reshape/reshape2 包.常用的有 melt 函数和 cast 函数.

melt 函数：根据数据类型选择 melt.data.frame、melt.array、 melt.list 函数进行实际操作。若是数组类型，melt 函数会依次对各维度的名称进行组合将数据进行向量化；若是列表数据，melt 函数将列表中的数据拉成两列，一列记录列表元素的值，另一列记录列表元素的名称。若列表中的元素是列表，则增加列变量存储元素名称。元素值排列在前，名称在后，越是顶级的列表元素名称越靠后。

cast 函数：可以用来还原数据、对数据进行汇总

（2）数据整形可以使用 plyr 包.

aaply、adply、alply 函数：将数组（array）分别转成数组、数据框和列表；daply、ddply、dlply 函数：将数据框分别转成数组、数据框和列表；laply、ldaply、llply 函数：将列表分别转成数组、数据框和列表。

（3）数据整形可以使用 apply 函数.

apply 函数：对数据进行循环、分组、过滤、类型控制等；

（4）数据整形可以使用 as.vector、un.list、transfrom、within、reshape、stack、unstack 函数.

as.vector 函数：将矩阵和多维数组向量化；

un.list 函数：将列表和数据框向量化；

transfrom、within 函数：为数据框增加新的列变量或改变列变量的值；

reshape、stack、unstack 函数：在数据框长格式和宽格式之间转换.

答 2：

常用汇总函数及描述统计的常用方法：

summary 函数: 对数据集做一般的描述,包括最值、分位数、均值等。当数据项是矩阵形式时将矩阵列看做向量.

min():求最小值；

max()：求最大值；

mean()：求均值同时当参数是表达式时候可用于计算频数可看做分位数的逆；

median()：求中值；sd()：求标准差；var()：求方差；

cor()：求相关系数；cov()：求协方差.

rnorm(n,mean=,sd=)生成服从正态分布的随机数；

runif(n,a,b)：生成 n 个 a 到 b 之间的均匀分布数.

答 3：

R 中常用的绘图函数及功能：

plot()：绘制两个变量的散点图、曲线图等；

pairs()：绘制多个变量的散点图，并以阵列形式排列；

pie()：绘制饼图；

parplot()：绘制条形图；

hist()：绘制直方图；

boxplot()：绘制箱线图；

qqnorm()：绘制 Q-Q 图.

2、读取 iris 数据集的满足特定条件的数据子集、作汇总表并使用描述统计常用函数并作结果解释.

答 1：

读取 iris 数据集的满足特定条件的数据子集、作汇总表：

代码如下：

```
# 取 iris 数据集的子集
iris.setosa<-iris[iris$Species=='setosa',]
iris.versicolor<-iris[iris$Species=='versicolor',]
iris.virginica<-iris[iris$Species=='virginica',]
# 作基本概要
summary(iris.setosa)
summary(iris.versicolor)
```

summary(iris.virginica)

结果如图 1 所示：

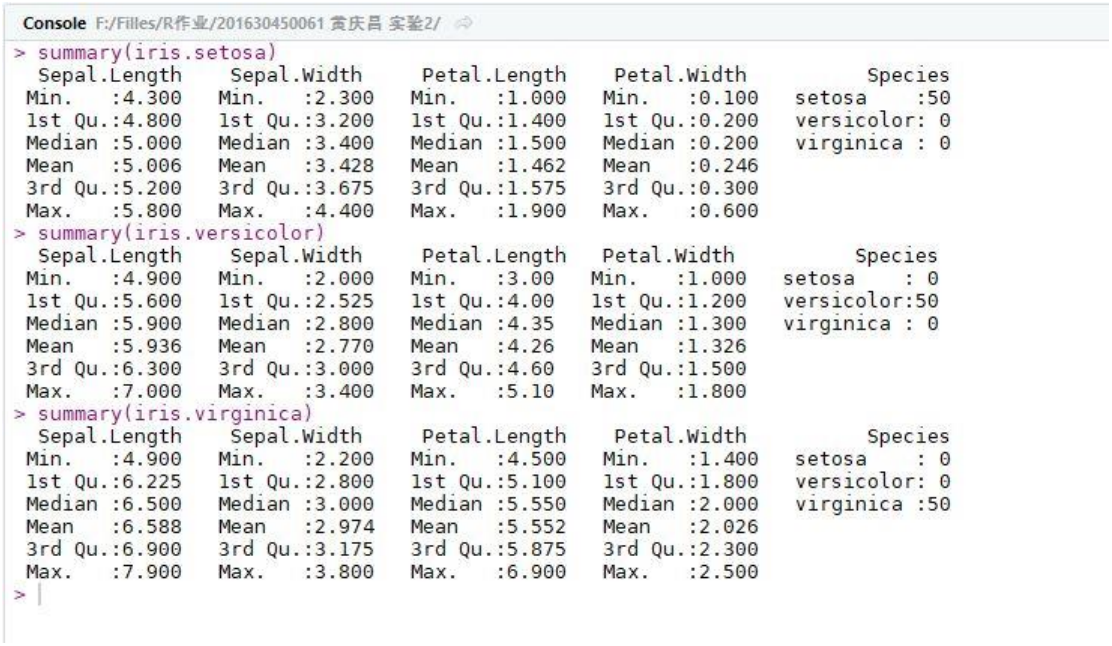


**图 1：iris 数据子集汇总**

答 2：

使用描述统计常用函数并作结果解释：

代码如下：

```
max(iris[1:4])
min(iris[1:4])
max(iris$Sepal.Length)
max(iris$Sepal.Width)
max(iris$Petal.Length)
max(iris$Petal.Width)
min(iris$Sepal.Length)
min(iris$Sepal.Width)
min(iris$Petal.Length)
min(iris$Petal.Width)
mean(iris$Sepal.Length)
mean(iris$Sepal.Width)
mean(iris$Petal.Length)
mean(iris$Petal.Width)
median(iris$Sepal.Length)
median(iris$Sepal.Width)
median(iris$Petal.Length)
median(iris$Petal.Width)
var(iris$Sepal.Length)
var(iris$Sepal.Width)
var(iris$Petal.Length)
```

```
var(iris$Petal.Width)
sd(iris$Sepal.Length)
sd(iris$Sepal.Width)
sd(iris$Petal.Length)
sd(iris$Petal.Width)
```

结果解释：求 iris 数据集各列 Sepal.Length、Sepal.Width、Petal.Length、Petal.Width 的最大值、最小值、均值、中位数、方差、标准差及前四列的最大值和最小值

3、作 iris 数据集的散点图，分变量直方图、箱线图、密度估计曲线

答：

（1）散点图绘制代码如下：

```
par(mfrow=c(2,2))
plot(iris$Sepal.Length,main='散点图 1')
plot(iris$Sepal.Width,main='散点图 2')
plot(iris$Petal.Length,main='散点图 3')
plot(iris$Petal.Width,main='散点图 4')
#绘制另外新的散点图
plot(iris[,1:4],main='散点图 5')
```

结果如图 2、图 3 所示：

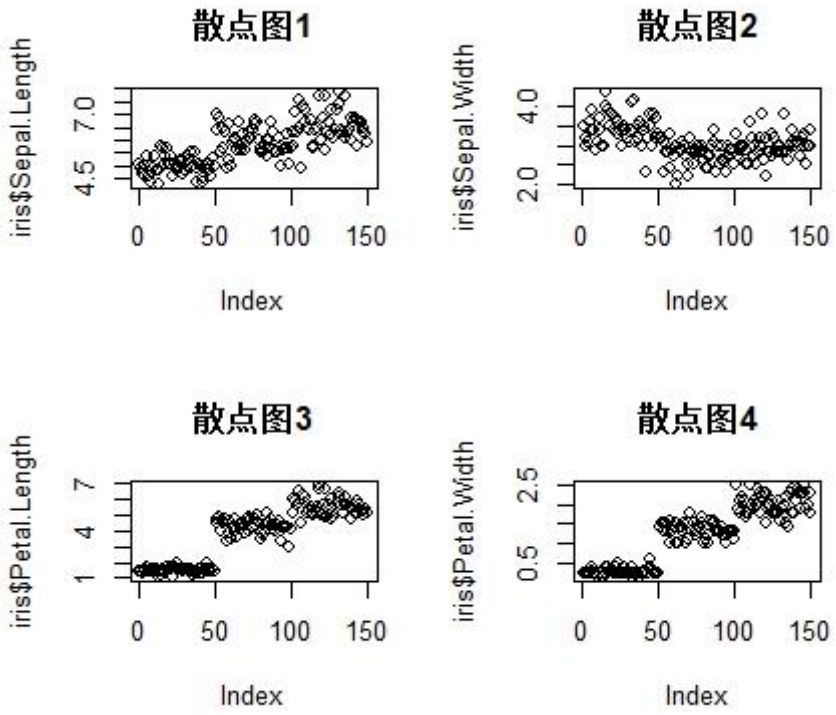


图 2：iris 数据集各属性的散点图

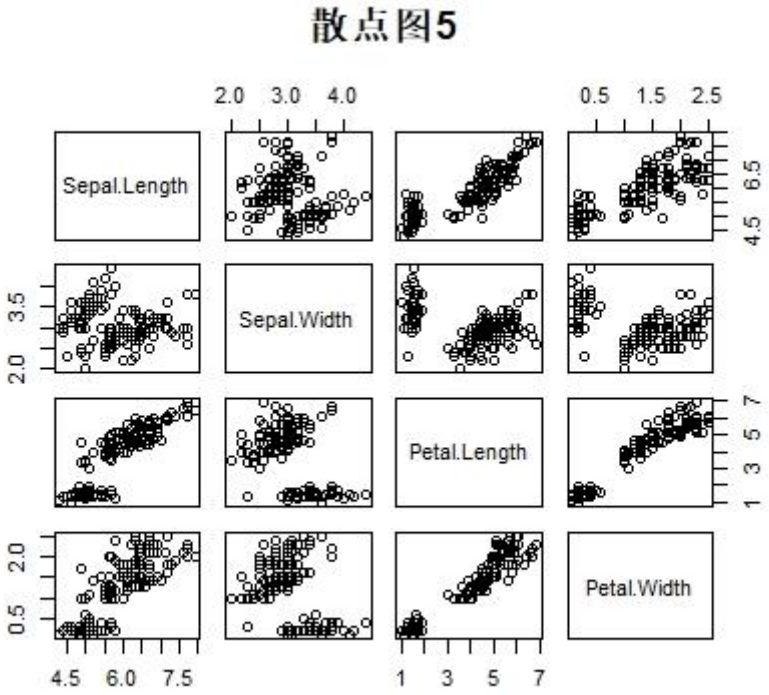


图 3：散点图 5

（2）各属性直方图如图 4 所示：

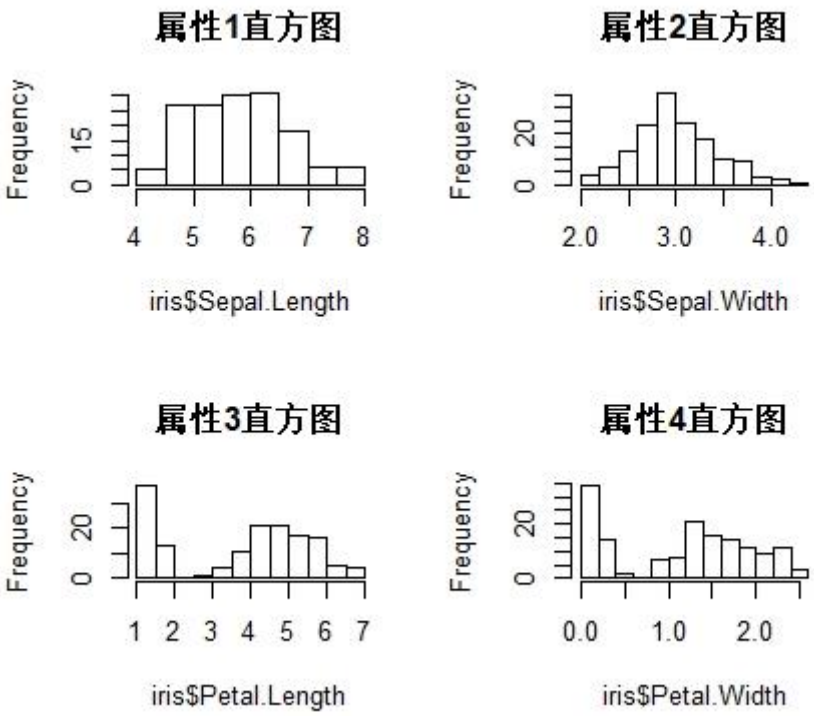


图 4：iris 数据集各属性直方图

代码如下所示：

```
par(mfrow=c(2,2))
hist(iris$Sepal.Length,main='属性 1 直方图')
```

```
hist(iris$Sepal.Width,main='属性 2 直方图')
hist(iris$Petal.Length,main='属性 3 直方图')
hist(iris$Petal.Width,main='属性 4 直方图')
```

（3）箱线图绘制代码如下所示：

```
par(mfrow=c(2,2))
boxplot(iris[,1],main='箱线图',ylab='Sepal.Length')
boxplot(iris[,2],main='箱线图',ylab='Sepal.Width')
boxplot(iris[,3],main='箱线图',ylab='Petal.Length')
boxplot(iris[,3],main='箱线图',ylab='Petal.width')
```

结果如图 5 所示：

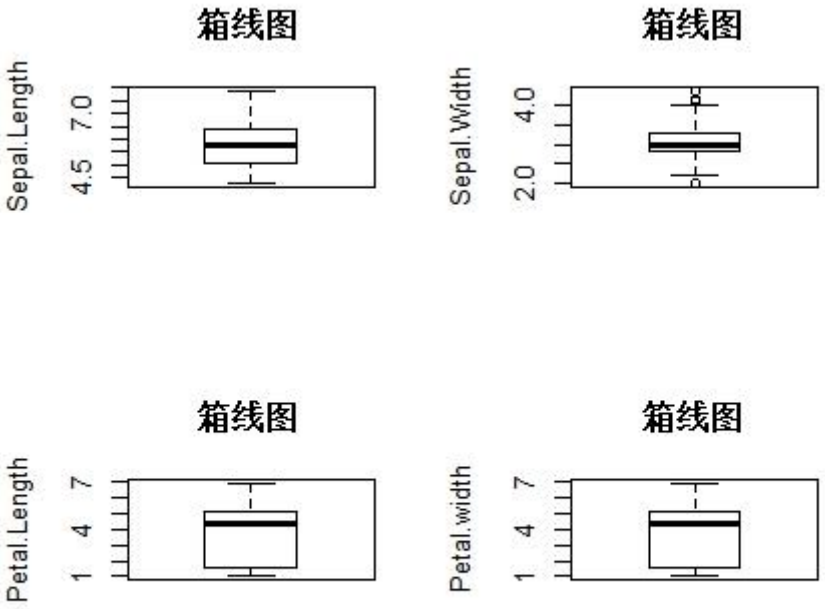


图 5：iris 数据集各属性箱线图

（4）密度估计曲线绘制代码如下：

```
d_Sepal.Length<-density(iris$Sepal.Length)
d_Sepal.Width<-density(iris$Sepal.Width)
d_Petal.Length<-density(iris$Petal.Length)
d_Petal.Width<-density(iris$Petal.Width)
par(mfrow=c(2,2))
plot(d_Sepal.Length,main='Sepal.Length 密度估计图',ylab='Density')
polygon(d_Sepal.Length,col='red')
plot(d_Sepal.Width,main='Sepal.Width 密度估计图',ylab='Density')
polygon(d_Sepal.Width,col='blue')
plot(d_Petal.Length,main='Petal.Length 密度估计图',ylab='Density')
polygon(d_Petal.Length,col='yellow')
plot(d_Petal.Width,main='Petal.Width 密度估计图',ylab='Density')
polygon(d_Petal.Width,col='green')
```

结果如图 6 所示：

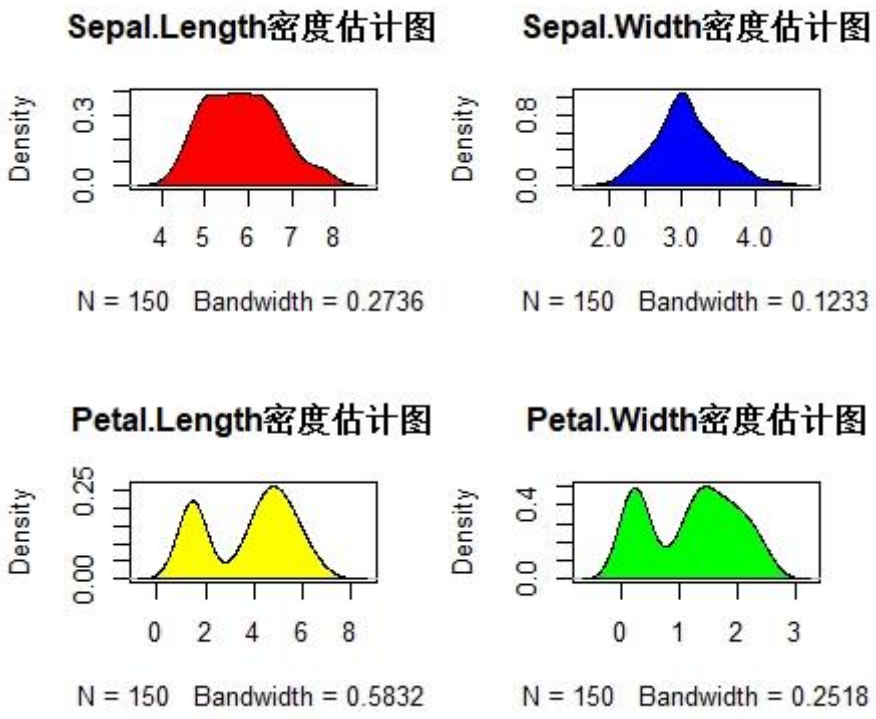


图 6：iris 各属性密度估计图

## 五、实验总结

通过这次实验，我认识了 R 中常用的汇总函数、统计功能函数、绘图函数及数据整形常用的库和函数，并通过对 R 中 iris 数据集数据整形操作及绘图，在实践中加深了理解和应用。

# 实验三 参数估计和假设检验的实现

| | | | |
|---|---|---|---|
| **地　　点：** | 4　楼　4105 房； | **实验台号：** | 个人电脑 |
| **实验日期与时间：** | 2019 年 5 月 10 日 | **评　　分：** | |
| **预习检查纪录：** | | **实验教师：** | 龙卫江 |
| **电子文档存放位置：** | C:\ex | | |
| **电子文档文件名：** | 201630450061 黄庆昌 实验三.doc | | |

**批改意见：**

# 实验三、参数估计和假设检验的实现

## 一、实验目的

1 理解参数估计的方法，掌握其 R 实现；能正确解读软件展示的结果的统计含义。

2 理解掌握假设检验的原理、方法及 R 实现；能正确解读软件展示的结果的统计含义。

3 熟悉 R 函数；自编函数实现一定统计分析功能。

## 二、实验环境

1 计算机（需联网）；2 R 系统及相关库；3 记录用的文具。

## 三、实验原理

参数估计和假设检验是经典统计的重要内容。原理上理解矩估计与大数定律的联系，最大似然的统计思想；理解区间估计和假设检验中小概率事件在联系统计与概率的作用，库诺桥。这些内容解释了本实验所涉及的统计方法的基本原理。要求学生掌握参数估计和假设检验的统计思想和统计软件主要的 R 函数。例如，置信区间的统计意义、假设检验中的库诺桥的使用。主要理解区间估计、假设检验以及能自编函数实现特定的统计功能，正确理解假设检验中的两类错误的意义，以及 p-值的含义和作用，并能准确地解释统计结果。

## 四、实验内容

1、简要叙述区间估计和假设检验的基本原理与主要步骤，常用公式。

答 1：

区间估计：设总体$X\sim F(x,\theta).\theta\in\vartheta$.如果统计量$T_1(X_1,\ldots,X_n),T_2(X_1,\ldots,X_n)$使得对给定的$\alpha\in(0,1)$，有$P(T_1\le g(\theta)\le T_2)=1-\alpha,\forall\theta\in\vartheta$，则称随机区间$[T_1,T_2]$为参数$g(\theta)$的置信度为$1-\alpha$的置信区间，$T_1,T_2$分别称为置信下界和置信上界.单个正态总体$X\sim N(\mu,\sigma^2)$的区间估计：

已知$\sigma^2$，$\mu$的置信区间为：$(\bar{X}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\bar{X}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$

未知$\sigma^2$，$\mu$的置信区间为：$(\bar{X}-t_{\alpha/2}\frac{S_{n-1}}{\sqrt{n}},\bar{X}+t_{\alpha/2}\frac{S_{n-1}}{\sqrt{n}})$

已知$\mu$，$\sigma^2$的置信区间为：$\left[\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{X^2_{1-\alpha/2}(n)},\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{X^2_{\alpha/2}(n)}\right]$

未知μ，$\sigma^2$的置信区间为：$\left[\frac{nS_n^2}{X_{1-\alpha/2}^2(n-1)},\frac{nS_n^2}{X_{\alpha/2}^2(n-1)}\right]$

答 2：

假设检验：设有来自某一参数分布族F(x,θ).θ ∈ ϑ其中θ为参数空间.原假设：$H_0:\theta\in\theta_0$，备选假设：$H_1:\theta\in H_1$，其中$\theta_1\in\vartheta$，$\theta_0\in\vartheta$，$\theta_0\cap\theta_1=\emptyset$假设检验中用到的统计量，称为检验统计量.检验统计量把样本空间分为两个区域，使$H_0$被拒绝的样本观察值所组成的区域称为拒绝域.此时，检验统计量落入拒绝域的概率是给定的小概率α，α称为显著性水平.

t 检验：若$X_1,\ldots,X_n$是来自正态总体X~N(μ,$\sigma^2$)的样本，且$\sigma^2$未知，则有：T = $\frac{\bar{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$，其中 $\bar{X}$ 为样本均值，S 为样本标准差.当$\sigma_1^2=\sigma_2^2=\sigma^2$时，

有：$$T=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)$$

F 检验：设$X_1,\ldots,X_{n1}$是来自正态总体X~N(μ,$\sigma_1^2$)的样本，$Y_1,\ldots,Y_{n2}$是来自正态总体Y~N(μ,$\sigma_2^2$)的样本,且两样本独立，则有：

$$F=\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}=\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2}\sim F(n_1-1,n_2-1)$$

2、简要叙述 R 基本包中假设检验函数的主要参数及使用，函数输出的解释和调用方式。

答 1：

在 R 中用 t.test()函数完成 t 检验的工作，并给出相应的置信区间，其使用格式为：

t.test(x,y=NULL,alternative=c("two.sided","less","greater"),mu=0,paired=FALSE,var.equal=FALSE,conf.level=0.95,…)

参数描述：

x、y为样本数值向量,alternative为备择假设选项；取"two.sided"（默认值）表示双侧检验，取"less"表示备择假设为"<"的单侧检验，取"greater"表示备择假设为">"的单侧检验；mu为数值表示原假设均值；paired为逻辑变量，用以说明是否完成配对数据的t检验；var.equal为逻辑变量，用以说明样本方差是否相同,conf.level为0-1之间的数值，表示置信水;formula 为公式,data

为矩阵或数据框.

答2：

在R中，用var.test()函数作F检验，其使用格式为：

var.test(x, y, ratio = 1, alternative = c(” two.sided”, ” less”, ” greater”),conf.level = 0.95, ...)

参数描述：

参数 x、y 分别为两个样本构成的数值向量；ratio 为两个总体的方差比，默认值为 1；alternative 为备择假设选项，取"two.sided"（默认值）表示双侧检验，取"less"表示备择假设为"<"的单侧检验，取"greater"表示备择假设为">"的单侧检验；conf.level 为 0-1 之间的数值，表示置信水平.

3、t.test()的使用：在 alpha 取为 0.06 时对 Fisher 鸢尾花数据中 setosa 花瓣长分别就原假设为 mu3=1.5， mu3=1.0， mu3>2 和 mu3<1.6 进行检验， 作出统计解释。

答 1：原假设为 mu3=1.5 时，代码如下：

```
a<-iris[iris$Species=='setosa',]
b<-a$Petal.Length
t.test(b,mu=1.5,conf.level = 0.94)
```

结果如图 1 所示：

```
> a<-iris[iris$Species=='setosa',]
> b<-a$Petal.Length
> t.test(b,mu=1.5,conf.level = 0.94)

 One Sample t-test

data:  b
t = -1.5472, df = 49, p-value = 0.1282
alternative hypothesis: true mean is not equal to 1.5
94 percent confidence interval:
 1.414714 1.509286
sample estimates:
mean of x
    1.462
```

**图 1：原假设 mu3=1.5 的 t 检验**

统计解释：因为 P 值（=0.1282）>0.06，不拒绝原假设，原假设成立.

答 2：原假设为 mu3=1.0 时，代码如下：

```
t.test(b,mu=1.0,conf.level = 0.94)
```

结果如图 2 所示：

```
> t.test(b,mu=1.0,conf.level = 0.94)

 One Sample t-test

data:  b
t = 18.811, df = 49, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 1
94 percent confidence interval:
 1.414714 1.509286
sample estimates:
mean of x
    1.462
```

图 2：原假设 mu3=1.0 的 t 检验

统计解释：因为 P 值（<2.2e-16）<0.06，拒绝原假设，即认为 mu3 不等于 1.0.

答 3：原假设为 mu3>2 时，代码如下：

t.test(b,mu=2,alternative = "less",conf.level = 0.94)

结果如图 3 所示：

```
> t.test(b,mu=2,alternative = "less",conf.level = 0.94)

 One Sample t-test

data:  b
t = -21.906, df = 49, p-value < 2.2e-16
alternative hypothesis: true mean is less than 2
94 percent confidence interval:
     -Inf 1.500863
sample estimates:
mean of x
    1.462
```

图 3：原假设 mu3>2 的 t 检验

统计解释：P 值（<2.2e-16）<0.06，拒绝原假设,即认为 mu3 不大于 2.

答 4：原假设为 mu3<1.6 时，代码如下：

t.test(b,mu=1.6,alternative = "greater",conf.level = 0.94)

结果如图 4 所示：

```
> t.test(b,mu=1.6,alternative = "greater",conf.level = 0.94)

 One Sample t-test

data:  b
t = -5.6189, df = 49, p-value = 1
alternative hypothesis: true mean is greater than 1.6
94 percent confidence interval:
 1.423137      Inf
sample estimates:
mean of x
    1.462
```

图 4：原假设 mu3<1.6 的 t 检验

统计解释：P 值（=1）>0.06，不拒绝原假设，即认为 mu3 小于 1.6.

4、对 x<- 1:30,在 alpha 取为 0.05 时,以 KS 检验为例检验其是否服从正态分布、指数分布和 Poisson 分布。发现什么问题？给出你的解释。

答 1：以 KS 检验为例检验其是否服从正态分布，代码如下：

x=1:30

ks.test(x,"pnorm",mean(x),sd(x))

结果如图 5 所示：

```
> x=1:30
> ks.test(x,"pnorm",mean(x),sd(x))

 One-sample Kolmogorov-Smirnov test

data:  x
D = 0.069849, p-value = 0.9963
alternative hypothesis: two-sided
```

图 5：KS 检验是否服从正态分布

统计解释：P 值（=0.9963）>0.05，不能拒绝原假设，即认为服从正态分布.

答 2：以 KS 检验为例检验其是否服从指数分布，代码如下：

ks.test(x,"pexp",1/mean(x))

结果如图 6 所示：

```
> ks.test(x,"pexp",1/mean(x))

 One-sample Kolmogorov-Smirnov test

data:  x
D = 0.17542, p-value = 0.2802
alternative hypothesis: two-sided
```

图 6：KS 检验是否服从指数分布

统计解释：P 值（=0.2802）>0.05，不能拒绝原假设，即认为服从指数分布.

答 3：以 KS 检验为例检验其是否服从 Poisson 分布，代码如下：

ks.test(x,"ppois",mean(x))

结果如图 7 所示：

```
> ks.test(x,"ppois",mean(x))

 One-sample Kolmogorov-Smirnov test

data:  x
D = 0.26376, p-value = 0.02485
alternative hypothesis: two-sided
```

图 7：KS 检验是否服从 Poisson 分布

统计解释：P 值（=0.02485）<0.05，拒绝原假设，即认为不服从 Poisson 分布.

5、自编 R 函数实现指定统计功能的函数：基本库中未提供的正态总体方差已知时的均值的区间估计及假设检验函数.

函数代码如下：

```
u.test<-function(x,mu=0,sigma,alpha=0.05,alternative="two.sided"){
  result<-list() # 初始化
  n<-length(x)
  mean<-mean(x) # 样本均值
```

```
  u<-(mean-mu)/(sigma*sqrt(1/n)) # u 检验统计量
  p<-pnorm(u,lower.tail = FALSE) # 计算 p 值
  qu<-qnorm(1-alpha/2,mean=0,sd=1,lower.tail = TRUE) # 计算分位数
  result$mean<-mean
  result$u<-u
  result$p.value<-p
  if(alternative=="two.sided"){
    p<-2*p
    result$p.value<-p
  }
  else if(alternative=="less"|alternative=="greater"){
    result$p.value<-p
  }
  else{
    return("input wrong")
  }
  result$conf.int<-c(mean-sigma*sqrt(1/n)*qu,mean+sigma*sqrt(1/n)*qu)
  result
}
```

## 五、实验总结

这次实验让我温习了数学上假设检验和参数估计的基本原理，学习了 R 中假设检验和参数估计常用的函数及其函数各参数的使用方法，并对 R 中 iris 数据集进行了相关操作，在实践中加深了自己的理解。

# 实验四 回归与方差分析的实现

| | | | |
|---|---|---|---|
| **地　　点：** | 4　楼　4105 房； | **实验台号：** | 个人电脑 |
| **实验日期与时间：** | 2019 年 5 月 17 日 | **评　　分：** | |
| **预习检查纪录：** | | **实验教师：** | 龙卫江 |
| **电子文档存放位置：** | C:\ex | | |
| **电子文档文件名：** | 201630450061 黄庆昌 实验四.doc | | |

**批改意见：**

# 实验四、回归与方差分析的实现

## 一、实验目的

1 理解掌握回归分析方法及 R 实现；理解回归诊断、岭回归、LASSO、CV、变量选择和模型选择等技术。

2 理解掌握方差分析方法及 R 实现。

## 二、实验环境

1 计算机（需联网）；2 R 系统及相关库；3 记录用的文具。

## 三、实验原理

回归和方差分析的实现是经典统计的重要内容，需要学生掌握其统计思想和主要的 R 函数，主要是对线性模型处理的技术，包括回归诊断、变量选择、模型选择、可视化展示等技术。最小二乘准则，偏差平方和的分解思想及解释。也理解均方差为代表的目标函数在经典统计的作用，偏差平方和分解原理和技术等。

## 四、实验内容

1、简要叙述回归分析的主要步骤，各步函数的调用和输出的解释。

（1）建立线性回归模型

设$X_1, X_2, \dots, X_p$为自变量，Y为因变量，且有$Y = \beta_0 + \beta_1 + \cdots + \beta_P X_p + \varepsilon$，其中$\varepsilon \sim N(0, \sigma^2), \beta_0, \beta_1, \dots, \beta_P$和$\sigma^2$是未知参数

（2）对回归系数进行估计

求最小二乘函数$Q(\beta) = (y - X\beta)^T (y - X\beta)$达到最小值的β

由计算可知：$\hat{\beta} = (X^T X)^{-1} X^T y$

（3）对回归系数进行显著性检验

回归系数的显著性检验是检验变量$X_j$的系数是否为 0，即假设检验为：

$$H_{j0}: \beta_j = 0, H_{j1}: \beta_j \neq 0, j = 0,1, \dots, p$$

当$H_{j0}$成立时，统计量

$$T_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{c_{jj}}} \sim t(n-p-1), j = 0,1,..., p$$

（4）回归方程的显著性检验

回归方程的显著性检验是检验方程的系数是否全为 0，即假设检验为：

$$H_0: \beta_0 = \beta_1 = \cdots = \beta_P = 0, H_1: \beta_0, \beta_1, \dots, \beta_P \text{不全为 } 0$$

当$H_0$成立时，统计量

$$\mathrm{F}=\frac{SSR/p}{SSE/(n-p-1)}\sim F(p,n-p-1)$$

当 F 统计量的 P 值<α时，拒绝原假设，即可以用线性方程来处理问题.

（5）相关性检验

在 R 中，lm()函数可以完成多元统计分析回归系数的估计、回归系数和回归方程的检验，其使用格式为：

lm(formula,data,subset,weights,na.action,method=qr”,model=TRUE,x=FALSE,y=FALSE,qr=TRUE,singular.ok=TRUE,constrasts=null,offest,...)

其中参数formula 为模型公式，subset 为可选项，表示所使用的子集，weights 为可选项向量，表示对样本的权重，na.action 为函数，表示当数据出现缺失的处理方法。method 为估计回归系数的方法

lm()函数需要用summary()函数提取结果.

2、简要叙述方差分析的主要原理和步骤, 函数的调用和输出的解释。

设试验只有一个因素 A 在变化,其它因素保持不变,A 有 r 个水平$A_1,A_2,\ldots,A_r$在水平$A_i$下进行$n_i$次独立观测，将水平$A_i$下的试验结果$x_{i1},x_{i2},\ldots,x_{in}$看成来自第 i 个正态总体$X_i\sim N(u_i,\sigma^2)$的样本观测值.其中$u_i$、$\sigma^2$均未知，并且每个总体$X_i$都相互独立.

主要原理：

单因素方差分析：比较因素 A 的 r 个水平的差异可以归结为比较这 r 个总体的均值。检验假设为：

$$H_0:\alpha_1=\alpha_2=\cdots=\alpha_r=0,H_1:\alpha_1,\alpha_2,\ldots,\alpha_r\text{不全为零}$$

若$H_0$被拒绝,则说明因素 A 的各水平的效应之间有显著的差异.当$H_0$成立时，采用

$$\mathrm{F}=\frac{S_A/(r-1)}{S_E/(n-r)}\sim\mathrm{F(r-1,n-r)}$$

其中$S_E=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x_i})^2$,$S_A=\sum_{i=1}^{r}(\bar{x_i}-\bar{x})^2$

作为$H_0$的检验统计量，计算 P 值，P 值小于α时拒绝原假设$H_0$

步骤：

（1）明确观测变量和控制变量；

（2）剖析观测变量的方差。方差分析认为：观测变量值的变动会受控制变量和随机变量两方面的影响。单因素方差分析将观测变量总的离差平方和分解为组间离差平方和和组内离差平方和两部分，用数学形式表述为：SST=SSA+SSE。

（3）通过比较观测变量总离差平方和各部分所占的比例，推断控制变量是否给观测变量带来了显著影响。

函数的调用：

aov (formula,data=NULL,projection=FALSE,qr=TRUE,constrasts=NULL,…)

其中，参数 formula 为公式，表示单因素或双因素方差分析，data 为数据框，表示数据与因素和水平的关系，默认为 NULL.

aov()函数需要用 summary()函数提取结果.

输出的解释：

在方差分析表中，Df 表示自由度，Sum Sq 表示平方和，Mean Sq 表示均方，F value 表示 F 值，即 F 比，Pr（>F）表示 P 值，A 是因素 A，Residuals 是残差，即误差.

3、由于地球绕太阳运转的轨道是一个椭圆，所以每年从地球上观测日轮的直径随着日期的不同而变化。每年年底地球离太阳的距离最近，因此观测到的直径也会大。下面给出每个月的第一天观测到的日轮直径的数据：

| 日期 | 1.1 | 2.1 | 3.1 | 4.1 | 5.1 | 6.1 | 7.1 | 8.1 | 9.1 | 10.1 | 11.1 | 12.1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 直径 | 32.6 | 32.5 | 32.3 | 32 | 31.8 | 31.6 | 31.5 | 31.6 | 31.7 | 32 | 32.3 | 32.5 |

试由上述数据建立预测日轮直径的数学模型。

e.g. 尝试作图猜测日轮直径关于时间（天）的近似可能关系，绝对值函数，三角函数 $Y = a + b\cos(\frac{2\pi t}{365}) + c\sin(\frac{2\pi t}{365}) + \varepsilon$ 等，写出 R 代码，解释运行结果，并尝试是否可能模型选择。

（1）画散点图知，日轮直径关于时间（天数）不为线性关系。故作线性模型：

$$Y = a + b\cos(\frac{2\pi t}{365}) + c\sin(\frac{2\pi t}{365}) + \varepsilon$$

代码如下：

```
t<-c(1,32,61,92,122,153,183,214,245,275,306,336)
d<-c(32.6,32.5,32.3,32,31.8,31.6,31.5,31.6,31.7,32,32.3,32.5)
plot(t,d)
s<-lm(d~cos(2*pi*t/365)+sin(2*pi*t/365))
summary(s)
```

结果如图 1 所示：

```
> t<-c(1,32,61,92,122,153,183,214,245,275,306,336)
> d<-c(32.6,32.5,32.3,32,31.8,31.6,31.5,31.6,31.7,32,32.3,32.5)
> plot(t,d)
> s<-lm(d~cos(2*pi*t/365)+sin(2*pi*t/365))
> summary(s)

Call:
lm(formula = d ~ cos(2 * pi * t/365) + sin(2 * pi * t/365))

Residuals:
      Min        1Q    Median        3Q       Max
-0.055895 -0.017864  0.005775  0.019614  0.039232

Coefficients:
                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)         32.030931   0.009593 3339.100  < 2e-16 ***
cos(2 * pi * t/365)  0.534359   0.013566   39.390 2.18e-11 ***
sin(2 * pi * t/365)  0.024137   0.013566    1.779    0.109
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03323 on 9 degrees of freedom
Multiple R-squared:  0.9942, Adjusted R-squared:  0.993
F-statistic: 777.4 on 2 and 9 DF,  p-value: 8.324e-11
```

图 1：关于天数的三角函数的线性回归模型

由图 1 知，P 值（=8.324e-11<0.05），拒绝原假设，即认为$\cos(2*\text{pi}*t/365)$、$\sin(2*\text{pi}*t/365)$与日轮直径存在线性关系，且回归方程为：

$$Y = 32.030931 + 0.534359*\cos\left(2*\text{pi}*\frac{t}{365}\right) + 0.024137*\sin(2*\text{pi}*\frac{t}{365})$$

4、利用不同配方的材料 A1,A2,A3 和 A4 生产出来的元件，抽样测得使用寿命数据如下。

A1：1600,1610,1650,1680,1700,1700,1780

A2：1500,1640,1400,1700,1750

A3：1640,1550,1600,1620,1640,1600,1740,1800

A4：1510,1520,1530,1570,1640,1600

据以往经验知$A_i$对应总体$\xi_i$服从正态分布,且满足方差齐性。取显著水平为 0.05。

1) 若假定总体仍服从正态分布并满足齐次性,检验四种配方材料使用寿命是否有显著差异?

2) 尝试检验是否仍可认为各总体$\xi_i$均服从正态分布，并具有方差齐性?(e.g. 使用 Shaprio 检验和 Bartlett 检验。)

答 1：

代码：

```
A1<-c(1600,1610,1650,1680,1700,1700,1780)
A2<-c(1500,1640,1400,1700,1750)
A3<-c(1640,1550,1600,1620,1640,1600,1740,1800)
A4<-c(1510,1520,1530,1570,1640,1600)
X<-c(A1,A2,A3,A4)
account=data.frame(X,A=factor(rep(1:4,c(7,5,8,6))))
a.aov<-aov(X~A,data = account) #判断四组数据是否具有先出差
summary(a.aov)
```

结果如图 2 所示：

```
> A1<-c(1600,1610,1650,1680,1700,1700,1780)
> A2<-c(1500,1640,1400,1700,1750)
> A3<-c(1640,1550,1600,1620,1640,1600,1740,1800)
> A4<-c(1510,1520,1530,1570,1640,1600)
> X<-c(A1,A2,A3,A4)
> account=data.frame(X,A=factor(rep(1:4,c(7,5,8,6))))
> a.aov<-aov(X~A,data = account) #判断四组数据是否具有先出差
> summary(a.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
A            3  49212   16404   2.166  0.121
Residuals   22 166622    7574
>
```

图 2：元件使用寿命方差分析

由图 2 知，P 值（0.121>0.05），故接受原假设，即认为这四种配方材料使用

寿命无显著差异.

答 2：

代码：

```
shapiro.test(A1) #正态性检验，判断各组数据是否正态
shapiro.test(A2)
shapiro.test(A3)
shapiro.test(A4)
bartlett.test(X~A,data = account)
```

结果如图 3 所示：

```
> shapiro.test(A1) #正态性检验，判断各组数据是否正态

 Shapiro-Wilk normality test

data:  A1
W = 0.94235, p-value = 0.6599

> shapiro.test(A2)

 Shapiro-Wilk normality test

data:  A2
W = 0.93842, p-value = 0.6548

> shapiro.test(A3)

 Shapiro-Wilk normality test

data:  A3
W = 0.88859, p-value = 0.2271

> shapiro.test(A4)

 Shapiro-Wilk normality test

data:  A4
W = 0.91768, p-value = 0.4888

> bartlett.test(X~A,data = account)  #判断是否具有方差齐性

 Bartlett test of homogeneity of variances

data:  X by A
Bartlett's K-squared = 5.8056, df = 3, p-value = 0.1215
```

图 3：Shaprio 检验和 Bartlett 检验

由图 3 知，各总体在 Shaprio 检验中的 P 值均>0.05, 可认为各总体 $\xi_i$ 均服从正态分布；由 Bartlett 检验结果知数据具有方差齐性

5、 失业率问题研究.

拟对研究某地区 1960 年到 1993 年的失业率进行研究。数据集 unemployment.csv 中含有该地区 34 个年份的观测数据，其中变量描述如下：

year:年份，unemp:失业率,，gdprate: 国民生产总值(GDP)的变动率，govspend:政府支出与 GDP 的比值，taxb:税收负担，salav:薪水与净增值的比值，infl:在 Philips 曲线中定义的通胀率。

将 unemployment.csv 读入数据框 unemployment。

a)考虑 unemp 作为 gdprate 的线性模型，进行回归分析,作显著性检验，给出回归方程，作图。

答 1：

代码：

```
unemployment<-read.csv('C:/Users/35088/Desktop/unemployment.csv')
x=unemployment$gdprate
y=unemployment$unemp
kkk<-lm(y~x)
summary(kkk)
plot(x,y,xlab = "gdprate",ylab = "unemp",col=4)
abline(10.0096,-1.3304)
```

结果如图 4 所示：

```
> unemployment<-read.csv('C:/Users/35088/Desktop/unemployment.csv')
> x=unemployment$gdprate
> y=unemployment$unemp
> kkk<-lm(y~x)
> summary(kkk)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3987 -1.5732 -0.2337  1.4000  5.6212

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.0996     0.8293   12.18 1.48e-13 ***
x            -1.3304     0.2069   -6.43 3.14e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.381 on 32 degrees of freedom
Multiple R-squared:  0.5637, Adjusted R-squared:  0.5501
F-statistic: 41.34 on 1 and 32 DF,  p-value: 3.144e-07
```

图 4：回归分析结果

由图 4 知，P 值（3.1444e-07<0.05），拒绝原假设，unemp 和 gdprate 存在线性关系且回归方程为：unemp=-1.3304*gdprate+10.0996.

画出散点图和回归方程，结果如图 5 所示：

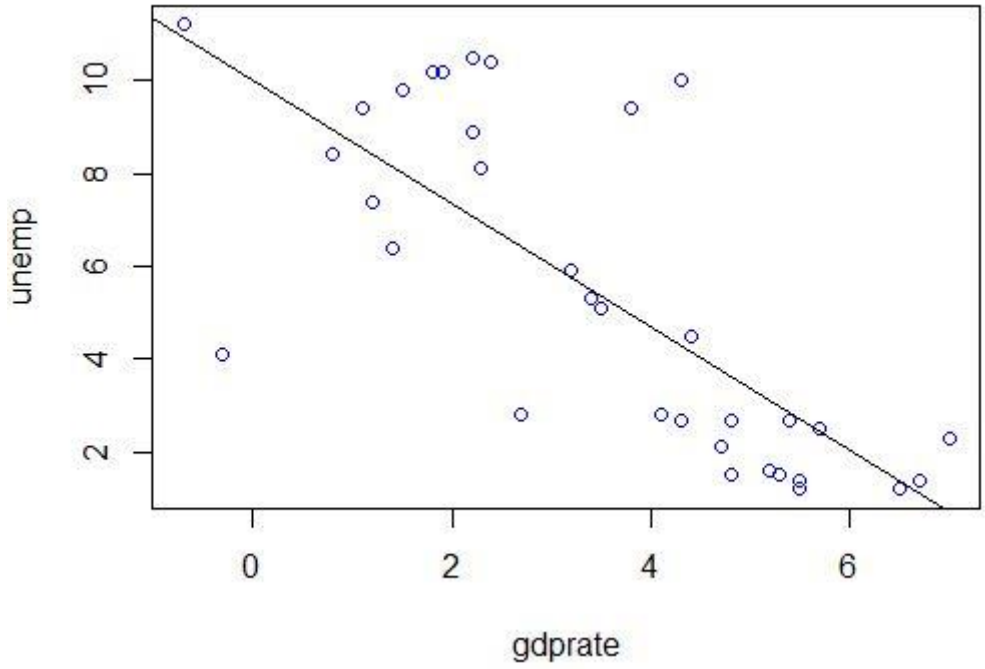


图 5：回归分析结果可视化

b)考虑 unemployment 去除 1993 年的数据，将 unemp 作为除 year 外的其他所有解释变量的多重线性模型建模，对结果解释。给出所有这些变量的相关矩阵，作出所有变量对的散点图。

答 2：

代码：

```
unemployment<-read.csv('C:/Users/35088/Desktop/unemployment.csv')
a<-unemployment[-34,]
ttt<-lm(a$unemp~1+a$gdprate+a$govspend+a$taxb+a$salav+a$infl)
summary(ttt)
cor(a[,-c(1,2)])
plot(a[,-c(1,2)],main='所有变量对的散点图')
```

结果如图 6 所示：

```
> unemployment<-read.csv('C:/Users/35088/Desktop/unemployment.csv')
> a<-unemployment[-34,]
> ttt<-lm(a$unemp~1+a$gdprate+a$govspend+a$taxb+a$salav+a$infl)
> summary(ttt)

Call:
lm(formula = a$unemp ~ 1 + a$gdprate + a$govspend + a$taxb +
    a$salav + a$infl)

Residuals:
     Min       1Q   Median       3Q      Max
-0.77821 -0.24058  0.03685  0.26530  1.34616

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.17741    4.15450  -1.728  0.09548 .
a$gdprate   -0.04519    0.08306  -0.544  0.59087
a$govspend   0.34070    0.12888   2.644  0.01350 *
a$taxb       0.25316    0.14637   1.730  0.09513 .
a$salav     -0.17922    0.05281  -3.394  0.00215 **
a$infl      -0.02847    0.03985  -0.714  0.48108
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4703 on 27 degrees of freedom
Multiple R-squared:  0.9844, Adjusted R-squared:  0.9815
F-statistic:   340 on 5 and 27 DF,  p-value: < 2.2e-16

> cor(a[,-c(1,2)])
            gdprate    govspend        taxb       salav       infl
gdprate   1.0000000 -0.77046669 -0.74706620 -0.25209391 -0.4079046
govspend -0.7704667  1.00000000  0.99214782  0.06875366  0.1141370
taxb     -0.7470662  0.99214782  1.00000000  0.02931499  0.1071257
salav    -0.2520939  0.06875366  0.02931499  1.00000000  0.6914845
infl     -0.4079046  0.11413696  0.10712567  0.69148445  1.0000000
```

图 6：多重线性模型

由图 6 知，P 值（<2.2e-16），拒绝原假设，即多重线性关系成立

模型的方程为：

unemp=-0.0452*gdprate+0.3407*govspend+02532*taxb-0.1792*salav-0.0285*infl-7.1774

各变量的相关矩阵如图 6 所示，这里采用的是默认的 Pearson 检验.

各变量的散点图如图 7 所示：

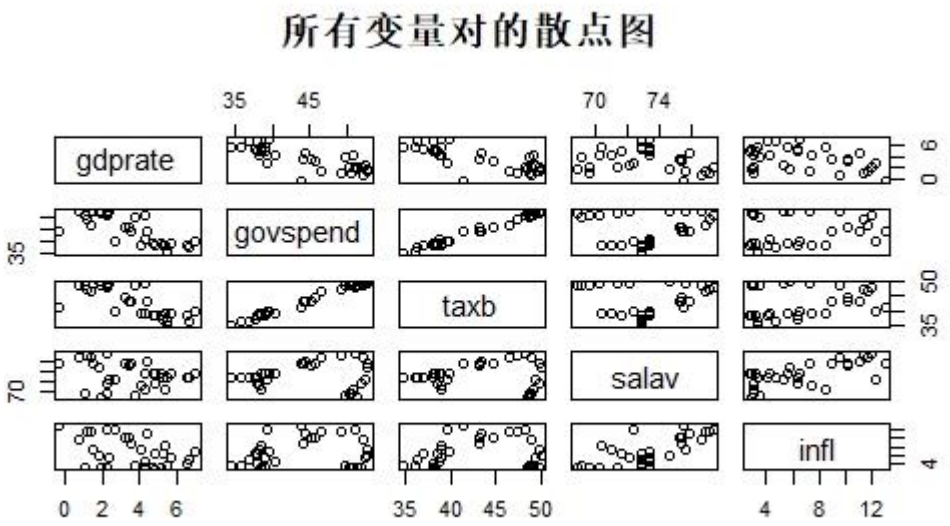


图 7：所有变量对的散点图

c)*以 b)中建模为基础，用 AIC 准则进行模型选择，给出最终模型。

代码：

```
lm.sol<-step(ttt)
summary(lm.sol)
```

结果如图 8 所示：

```
> lm.sol<-step(ttt)
Start:  AIC=-44.41
a$unemp ~ 1 + a$gdprate + a$govspend + a$taxb + a$salav + a$infl

             Df Sum of Sq    RSS     AIC
- a$gdprate   1   0.06548 6.0385 -46.046
- a$infl      1   0.11292 6.0859 -45.788
<none>                    5.9730 -44.406
- a$taxb      1   0.66177 6.6347 -42.938
- a$govspend  1   1.54593 7.5189 -38.810
- a$salav     1   2.54757 8.5205 -34.683

Step:  AIC=-46.05
a$unemp ~ a$govspend + a$taxb + a$salav + a$infl

             Df Sum of Sq    RSS     AIC
- a$infl      1   0.05971 6.0982 -47.721
<none>                    6.0385 -46.046
- a$taxb      1   0.59795 6.6364 -44.930
- a$govspend  1   2.15084 8.1893 -37.991
- a$salav     1   2.74755 8.7860 -35.670

Step:  AIC=-47.72
a$unemp ~ a$govspend + a$taxb + a$salav

             Df Sum of Sq     RSS     AIC
<none>                     6.0982 -47.721
- a$taxb      1    0.5418  6.6400 -46.912
- a$govspend  1    2.4464  8.5445 -38.590
- a$salav     1    6.7750 12.8732 -25.065
```

图 8：用 AIC 准则进行模型选择

去掉变量 gdprate、infl 能使 AIC 达到最小，其效果如图 9 所示：

```
> summary(lm.sol)

Call:
lm(formula = a$unemp ~ a$govspend + a$taxb + a$salav)

Residuals:
     Min       1Q   Median       3Q      Max
-0.74124 -0.24126 -0.00223  0.23244  1.42420

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.96419    3.01019  -1.981  0.05711 .
a$govspend   0.38269    0.11220   3.411  0.00193 **
a$taxb       0.21429    0.13350   1.605  0.11929
a$salav     -0.20302    0.03577  -5.676 3.89e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4586 on 29 degrees of freedom
Multiple R-squared:  0.984, Adjusted R-squared:  0.9824
F-statistic: 595.9 on 3 and 29 DF,  p-value: < 2.2e-16
```

图 9：逐步回归效果

由图知，回归系数的显著性水平检验有很大的提高，回归方程为：

unemp=0.3827*govspend+0.2143*taxb-0.2030*salav-5.9642

d)在你给出的模型里，已知 1993 年的解释变量值下求失业率的 0.95 预测区间。它包含了 1993 年失业率的观测值吗？

代码：

```
unemployment<-read.csv('C:/Users/35088/Desktop/unemployment.csv')
a<-unemployment[-34,]
x<-a$govspend
y<-a$taxb
z<-a$salav
q<-lm(a$unemp~x+y+z)
predict(q,data.frame(x=52.2,y=49.0,z=68.6),interval = "confidence")
```

结果如图 10 所示：

```
> unemployment<-read.csv('C:/Users/35088/Desktop/unemployment.csv')
> a<-unemployment[-34,]
> x<-a$govspend
> y<-a$taxb
> z<-a$salav
> q<-lm(a$unemp~x+y+z)
> predict(q,data.frame(x=52.2,y=49.0,z=68.6),interval = "confidence")
       fit      lwr      upr
1 10.58529 10.04432 11.12626
>
```

图 10:求预测值置信区间

由图 10 知，预测值为 10.58529，预测值的置信区间为[10.04432,11.12626]而 1993 年失业率的观测值为 11.2，故没有包含 1993 年失业率的观测值

e)*尝试用岭回归和 LASSO 模型建模。

代码 1：

```
a<-read.csv('C:/Users/35088/Desktop/unemployment.csv')
b<-a[,-1]
library(MASS)
rid1<-lm.ridge(unemp~gdprate+govspend+taxb+salav+infl,data=b,lambda = seq(0,0.5,0.001))
plot(rid1)
select(rid1)
rid2<-lm.ridge(unemp~gdprate+govspend+taxb+salav+infl,data=a,lambda = seq(0.105))
rid3<-lm.ridge(unemp~gdprate+govspend+taxb+salav+infl,data=a,lambda = seq(0.055))
library(ridge)
mod <-linearRidge(unemp~.,data =b)
summary(mod)
```

结果如图 11 所示：

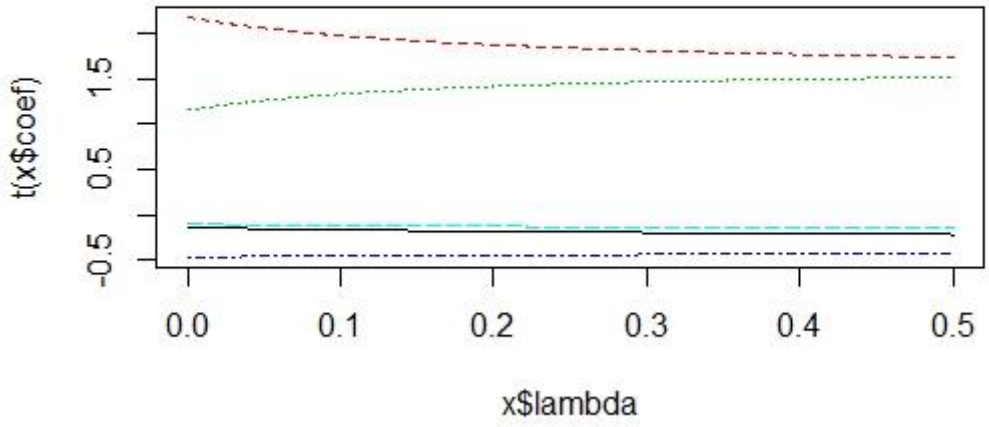


图 11：岭迹曲线

```
> a<-read.csv('C:/Users/35088/Desktop/unemployment.csv')
> b<-a[,-1]
> library(MASS)
> rid1<-lm.ridge(unemp~gdprate+govspend+taxb+salav+infl,data=b,lambda = seq(0,0.5,0.001))
> plot(rid1)
> select(rid1)
modified HKB estimator is 0.1046308
modified L-W estimator is 0.05473872
smallest value of GCV  at 0.5
```

图 12

```
> library(ridge)
> mod <-linearRidge(unemp~.,data =b)
> summary(mod)

Call:
linearRidge(formula = unemp ~ ., data = b)


Coefficients:
            Estimate Scaled estimate Std. Error (scaled) t value (scaled) Pr(>|t|)
(Intercept) -6.40303              NA                  NA               NA       NA
gdprate     -0.09445        -1.08711             0.78914            1.378    0.168
govspend     0.30727        10.86794             2.15361            5.046 4.50e-07 ***
taxb         0.28067         8.27734             2.03922            4.059 4.93e-05 ***
salav       -0.18229        -2.59910             0.65520            3.967 7.28e-05 ***
infl        -0.03915        -0.76339             0.69813            1.093    0.274
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge parameter: 0.006156898, chosen automatically, computed using 3 PCs

Degrees of freedom: model 4.432 , variance 4.132 , residual 4.733
```

图 13

测岭回归参数值为 0.00615,各自变量的系数显著性明显提高

代码 2：

```
library(lars)
x=as.matrix(b[,2:5])
y=as.matrix(b[,1])
(laa=lars(x,y,type = "lar"))
plot(laa)
```

结果如图 14 所示：

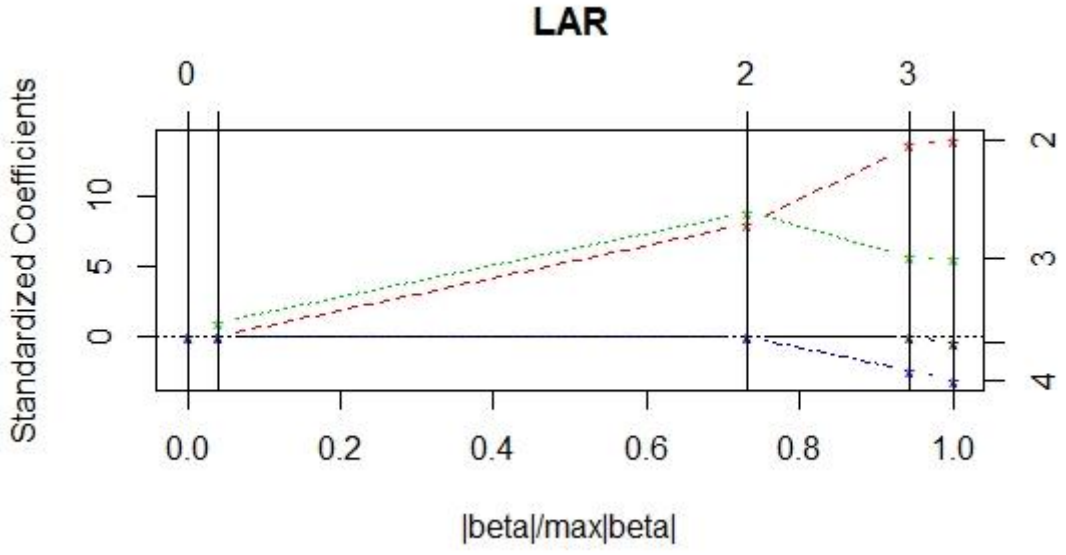


图 14

由此可见，LASSO 的变量选择依次是 govspend、taxb、salav、gdprate

```
> y = as.matrix(b[,1])
> (laa = lars(x, y, type = "lar"))

Call:
lars(x = x, y = y, type = "lar")
R-squared: 0.985
Sequence of LAR moves:
     taxb govspend salav gdprate
Var     3        2     4       1
Step    1        2     3       4
> plot(laa)
> summary(laa)
LARS/LAR
Call: lars(x = x, y = y, type = "lar")
  Df    Rss        Cp
0  1 415.96 1884.1819
1  2 380.20 1721.4268
2  3  24.90   86.7253
3  4   7.04    6.4425
4  5   6.30    5.0000
>
```

图 15

## 五、实验总结

通过这次实验，我学习了线性回归和方差分析的有关理论，也学会了在线性回归效果不好时，对数据进行逐步回归、岭回归等，并通过在 R 中的实际操作，在实践中加深了理解和应用。

# 实验五 多元统计分析的技术及实现

| | | | |
|---|---|---|---|
| **地　　点：** | 4　楼　4105 房； | **实验台号：** | 个人电脑 |
| **实验日期与时间：** | 2019 年 5 月 24 日 | **评　　分：** | |
| **预习检查纪录：** | | **实验教师：** | 龙卫江 |
| **电子文档存放位置：** | C:\ex | | |
| **电子文档文件名：** | 201630450061 黄庆昌 实验五.doc | | |

**批改意见**

# 实验五、多元统计分析的技术及实现

## 一、实验目的

1、理解多元分析中的降维技术，掌握其 R 实现；

2、理解多元统计中主要的监督学习方法，掌握其 R 实现；

3、解多元统计中主要的非监督学习方法及 R 实现。

## 二、实验环境

1 计算机（需联网）；2 R 系统及相关库；3 记录用的文具。

## 三、实验原理

多元统计是统计学的重要内容，理论涉及知识广泛，应用遍及社会、工程、技术、科学各个领域，经典多元统计分析方法的导出多以分布、距离、相似性等为基础，构造目标函数并借助优化工具解决问题。需要学生掌握多元统计分析的思想和常用的 R 函数，主要选择了降维技术中的 PCA、FA, 监督学习中的判别分析，非监督学习的聚类分析，这些内容是从数据中学习的重要的工具。

## 四、实验内容

1、简要叙述 PCA 方法，R 函数调用及输出的解释。

设$\mathrm{X}=(X_1,X_2,\ldots,X_p)^T$为 p 维随机变量，考虑线性变换:

$Z_1=a_{11}X_1+a_{21}X_2+\cdots+a_{p1}X_p$

$Z_2=a_{12}X_1+a_{22}X_2+\cdots+a_{p2}X_p$

…

$Z_p=a_{1p}X_1+a_{2p}X_2+\cdots+a_{pp}X_p$

并且满足

（1）$Z_i,Z_j$相互独立；

（2）$\mathrm{var}(Z_1)\geq var(Z_2)\geq\cdots\geq var(Z_p)$;

（3）$a_{1k}^2+a_{2k}^2+\cdots+a_{pk}^2=1,k=1,2,\ldots,p.$

称$Z_k$为原始变量$X_1,X_2,\ldots,X_p$的第 k 个主成分

R 函数调用：

1、princomp 函数

（1）公式形式：

princomp(formula,data=NULL,subset,na.action)

参数描述：

formula为公式，无响应变量，data 为数据框,subset 为可选向量, 表示

选择的样本子集，na.action 为函数，表示缺失数据的处理方法。

矩阵形式：

（2）矩阵形式

princomp(x,cor=FALSE,scores=TRUE,covmat=NULL,subset=rep(TRUE,nrow(as.matrix(x))))

参数描述：

x为数据框或数值矩阵，即用于主成分分析的样本，cor为逻辑变量，取TRUE表示用样本的相关矩阵作主成分分析,scores为逻辑变量，表示是否计算各主成分的分量,covmat 为协方差矩阵.

2、loading函数

loading()函数是显示主成分分析中的载荷矩阵.

使用格式为loading(x),其中参数x为princomp()函数或factanal()函数生成的对象。

3、predict函数

predict()函数用于计算主成分得分，其使用格式为：

predict(object,newdata,…)，其中参数object为princomp()函数生成的对象，newdata为预测值构成的数据框，默认值为全体观测样本。

4、screeplot函数

screeplot()函数的主要功能是画出主成分的碎石图，其使用格式为：

screeplot(x,npcs=min(10,length(x$sdev)),type=c(“barplot”,”lines”),main=deparse(substitute(x)),…)其中参数x为包含标准差的对象，npcs为画出的主成分的个数，type为字符串，描述所画碎石图的类型。

5、biplot函数

使用格式：biplot(x，choices=1:2,scale=1,pc.biplot=FALSE)

参数描述：x为princomp() 生成的对象，choices为二位数值向量，表示选择低级主成分，默认为第一和第二主成分，scale为0到1之间的数值，默认为1,pc.biplot为逻辑变量，取TRUE表示使用GABRIEL提出的画图方法，默认为FALSE

2、简要判别分析主要原理和步骤, 函数的调用和输出的解释。

主要原理：

设有 k 个不同的 p 维总体$G_1,G_2,\dots,G_k$,它们的分布函数为$F_1(y),F_2(y),\dots,F_k(y)$.现有一个属于这 k 个总体之一的样本，利用训练样本，将 p 维空间划分为 k 个互不相干的区域$R_1,R_2,\dots,R_k$,当$\mathrm{x}\in R_i$时，就判定 x 属于总体$G_i(i=1,2,\dots,k)$.

R 函数调用：

（1）线性判别函数lda()和二次判别函数qda():

lda(x,grouping,prior=proportions,tol=1:0e4,method,CV=FALSE,nu,……)

qda(x,grouping,prior=proportions,tol=1:0e4,method,CV=FALSE,nu,……)

lda(formula,data,…,subset,na.action)

qda(formula,data,…,subset,na.action)

参数解释：

x为矩阵或数据框，或者包含解释变量的矩阵,grouping 是指定样本属于哪一类的因子向量，prior 是各类的先验概率，默认是已有训练样本的计算结果，tol 控制精度，用于判别矩阵是否奇异,method 是估计方法（mle: 极大似然估计mve: 使用cov.mve 做估计,t: 基于t 分布的稳健估计）,CV 返回值包含留一法交叉确认情况,nu 是method=”t”的自由度。

输出解释：

lda() 的返回值有：调用方法，先验概率，每一类样本的均值，和线性判别系数。qda() 返回值没有现行判别系数其他与lda() 相同。

（2）预测函数 predict()

线性：predict(object,ndata,prior=object$prior,dimen,method,…)

二次：predict(object,ndata,prior=object$prior,method,…)

参数描述：

object 为判别函数（lda，qda）生成的对象,ndata 是预测数据构成的数据框（判别函数使用公式计算）或者向量。默认为全体样本；prior 为先验概率,dimen 使用空间的维数,method 参数估计的方法。

3、简要聚类分析主要原理和步骤, 函数的调用和输出的解释。

主要原理：

聚类分析是根据在数据中发现的描述对象及其关系的信息,将数据对象分组。目的是使组内的对象相互之间是相关的，而不同组中的对象是不相关的。组内相似性越大，组间差距越大，说明聚类效果越好。

步骤：

（1）定义问题与选择分类变量；

（2）选择聚类方法；

（3）确定群组数目；

（4）对聚类结果进行评估；

（5）结果的描述、解释.

R 函数调用：

（1）距离函数: dist(x,method=”euclidean”,diag=FALSE,upper=FALSE,p==)

参数解释：

x 为数值矩阵、数据框或是”dist”的对象，method 为定义距离的方法，diag 为逻辑变量，表示是否输出对角线上的距离，默认值为 FALSE，upper 为逻辑变量，表示输出上三角阵的值，默认值为 FALSE,表示仅输出下三角矩阵的值.p 为 Minkowski 距离的参数 q,默认值为 2，即欧氏距离。

（2）数据变换：scale(x,center=TRUE,scale=TRUE)

参数解释：

x 为样本构成的数值矩阵，center 为逻辑变量，表示是否对数据作中心变换，默认值为 TRUE；或者为数值向量，其维数等于矩阵 x 的列数，表示以 center 为中心作变换，scale 为逻辑变量，表示是否对数据作标准变换，默认值为 TRUE；或者为数值向量，其维数等于矩阵 x 的列数，表示以 scale 为尺度作标准变换.

（3）系统聚类：hclust(d,method=”complte”,members=NULL)

参数解释：

d 为 dist()函数生成的对象，method 为系统聚类的方法，merbers 或者为 NULL（默认值），或者为与 d 有相同变量长度的向量.

4、现收集了某班 52 位学生的数学、物理、化学、语文、历史和英语课程的考试成绩(X1-X6)，形成数据文件 StuSco.xls。尝试就学生成绩数据进行分析。

4.1)读入 StuSco.xls 到数据框 stusco， 尝试做主成分分析。a)求样本相关系数矩阵，保留三位小数。b)使用样本相关系数阵做主成分分析。c)列出主成分分析分析结果，写出主成分，并尝试对前两主成分给出较合理的解释。d)画出碎石图。e)计算主成分得分。f)提取主成分载荷矩阵。g)绘制主成分散点图。

4.2)*尝试 stusco 做因子分析,对运行结果给出恰当解释,并给出可视化展示。

答 1：

代码：

```
library(xlsx)
stusco<-read.xlsx("C:/Users/35088/Desktop/StuSco.xls",1)
round(cor(stusco),3)
pr<-princomp(stusco,cor = TRUE)
summary(pr,loadings = TRUE)
screeplot(pr,type = "lines")
predict(pr)
loadings(pr)
biplot(pr,scale = 0.5)
```

a)样本相关系数矩阵如图 1 所示：

```
> library(xlsx)
> stusco<-read.xlsx("C:/Users/35088/Desktop/StuSco.xls",1)
> round(cor(stusco),3)
       x1     x2     x3     x4     x5     x6
x1  1.000  0.647  0.696 -0.561 -0.456 -0.439
x2  0.647  1.000  0.573 -0.503 -0.351 -0.458
x3  0.696  0.573  1.000 -0.380 -0.274 -0.244
x4 -0.561 -0.503 -0.380  1.000  0.813  0.835
x5 -0.456 -0.351 -0.274  0.813  1.000  0.819
x6 -0.439 -0.458 -0.244  0.835  0.819  1.000
```

图 1：样本相关系数矩阵

b、c)主成分分析结果如图 2 所示：

```
> pr<-princomp(stusco,cor = TRUE)
> summary(pr,loadings = TRUE)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
Standard deviation     1.9261112 1.1236019 0.66395522 0.52009785 0.41172308 0.38309295
Proportion of Variance 0.6183174 0.2104135 0.07347275 0.04508363 0.02825265 0.02446003
Cumulative Proportion  0.6183174 0.8287309 0.90220369 0.94728732 0.97553997 1.00000000

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
x1  0.412  0.376  0.216  0.788         0.145
x2  0.381  0.357 -0.806 -0.118 -0.212 -0.141
x3  0.332  0.563  0.467 -0.588
x4 -0.461  0.279               -0.599  0.590
x5 -0.421  0.415 -0.250         0.738  0.205
x6 -0.430  0.407  0.146  0.134 -0.222 -0.749
```

图 2：主成分分析结果

由于前 4 个主成分的累积贡献率已达到 94.72%，另外两个主成分可以舍去，达到降维的目的。Comp.1、Comp.2、Comp.3、Comp.4 与原始变量的线性关系为：

$$\text{Comp.}1 = 0.412x_1 + 0.381x_2 + 0.332x_3 - 0.461x_4 - 0.421x_5 - 0.430x_6$$

$$\text{Comp.}2 = 0.376x_1 + 0.357x_2 + 0.563x_3 + 0.279x_4 + 0.415x_5 + 0.407x_6$$

$$\text{Comp.}3 = 0.216x_1 - 0.806x_2 + 0.467x_3 - 0.250x_5 - 0.146x_6$$

$$\text{Comp.}4 = 0.788x_1 - 0.118x_2 - 0.588x_3 + 0.134x_6$$

第一个主成分 Comp.1 是一定比例数学、物理、化学成绩与语文、历史、英语成绩相减，它的大小和正负反映了学生在学习上是偏向文科（语文、历史、英语）或是理科（数学、物理、化学）；第二个主成分 Comp.2 对应系数的符号都相同，它反映了学生的综合学习能力，学习能力越好的学生，Comp.2 数值越高。

d、e)主成分的碎石图如图 3 所示，主成分得分如图 4 所示：

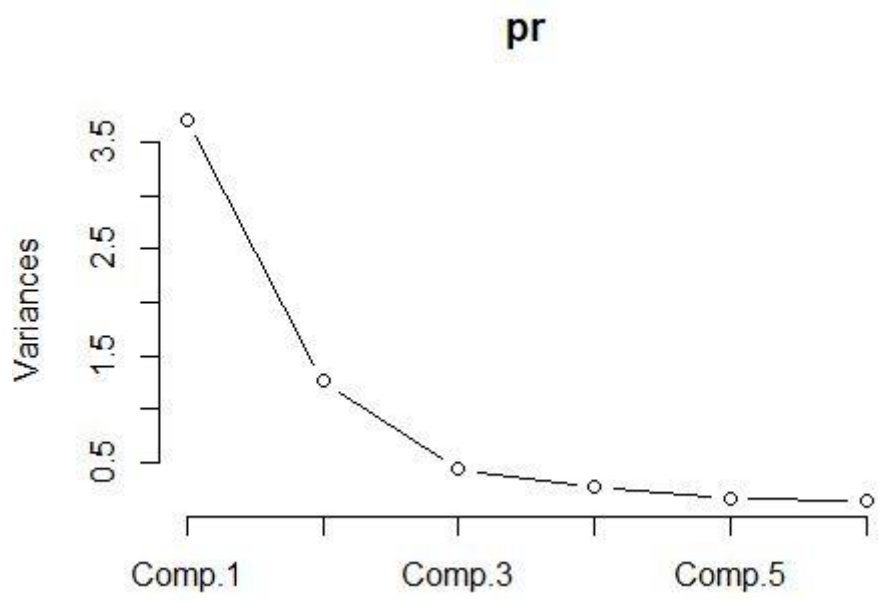


图 3：主成分碎石图

```
> predict(pr)
           Comp.1      Comp.2      Comp.3      Comp.4      Comp.5      Comp.6
 [1,] -1.84585050 -0.09869937  0.501399041 -0.37898025 -0.218954549  0.19042313
 [2,]  1.38293746 -0.93264526 -0.027536570 -0.16148676  0.146142481  0.75077808
 [3,] -0.04447490 -2.80445313 -0.220442506  0.36780857 -0.047954274  0.41604476
 [4,]  1.25557209 -0.40584893 -0.350765626  0.01079265  0.020616784 -0.02503406
 [5,]  1.13902555 -2.26874265 -0.004069094 -0.58347250  0.325721469 -0.21181294
 [6,]  3.51753434 -0.81974747 -1.071937337 -0.15576301  0.762687393 -0.16611292
 [7,]  3.51620268  0.10402783  0.100691576 -0.57359813  0.010606219  0.07976680
 [8,]  0.98230531 -2.32640027 -1.291921517 -0.01669947 -0.092817627 -0.05206543
 [9,] -2.24663266  0.12983805  0.407592509  0.22352473  1.289305161  0.06013990
[10,]  1.67501924 -0.32466680 -0.499446161 -0.38712414 -0.658485749  0.38595135
[11,]  1.17695408 -0.76065098  0.802575390  0.59520424  0.334402720  0.19076785
[12,] -1.15656395  0.21478424  0.367515379 -1.16996433  0.682285235 -0.21704343
[13,]  1.94424108  0.78320441  1.290177317  0.17426050  0.213864401  0.26356014
[14,] -0.13026004  0.90867535  0.944082148  0.39720659  0.015121216 -0.85208335
[15,]  0.08940481 -1.27591036  0.776266373 -0.78466689  0.518868579  0.10211225
[16,]  1.62140217  0.05220103  0.369315264  0.87300205  0.324476384 -0.03511783
[17,]  0.65077543  0.54509667 -0.560713720 -0.20766918 -0.377491545  0.11646716
[18,] -1.70890651  1.01214990  0.110970297 -0.08135398  0.616401108  0.44398269
[19,] -2.81890001  0.87577409 -0.280826182  0.74593951 -0.022449694  0.12871355
[20,] -0.04479664  1.50659205 -0.700307316  0.29526373  0.193652165 -0.33540679
[21,] -1.74763780  1.18458088 -1.044624100 -0.15401834 -0.106325395 -0.35083207
[22,]  1.87131409  1.17348168 -0.048992458 -0.32504480  0.280411924  0.14685910
[23,]  1.89649934  1.48958061  1.011850823  0.07386318 -1.015054543 -0.29993615
[24,]  0.45935156  0.12565877  0.295795299 -0.49902127  0.845265017 -0.12236024
[25,]  2.73741142  0.57906315 -0.676849729  0.20496123  0.388196122 -0.73266727
[26,]  0.84129968  2.11740575  0.543837540 -0.15613980  0.070441461  0.19228409
[27,] -0.70743714 -0.66945216 -0.751948001 -0.37697967 -0.244000194 -0.37653575
[28,] -1.95932966 -1.59387782  0.748826610 -0.23763055 -0.309220363 -0.07554685
[29,] -0.97691254 -0.03447987 -0.116644873 -1.43833992 -0.523222558  0.26620406
[30,] -4.48954778  0.69279938 -0.619582483  0.83234140  0.054347974  0.02883872
[31,] -2.95811586 -1.11323328  0.693204266 -0.46458327 -0.049942671  0.12386520
[32,] -2.21363704 -1.07153767  1.413998473 -0.24885858 -0.162669191 -0.32166138
[33,] -0.34542299  2.18736513  0.413742382  0.22463160 -0.018438019  0.85424952
[34,]  2.21123379  0.12884205 -0.690804852 -0.30127090 -0.423296645 -0.52749199
[35,]  0.56119163  0.31668122  0.688799889  0.38678380  0.486195952 -0.24200845
[36,]  0.35593895 -1.13107716 -1.018716128 -0.41947012  0.027967471  0.39509239
[37,] -1.88151345 -1.49590296 -0.086154549  1.16687824 -0.368740791  0.41912805
[38,]  2.93931631 -0.10994465  0.743891991  0.13900272 -0.477488575  0.18616233
[39,] -0.77304507  0.08283856 -0.685503726  0.37344659 -0.151147788 -0.31741632
[40,]  1.76336851  1.22435661 -0.793289044  0.10205140  0.119668330  0.46871314
[41,] -1.19078307  0.54556585  0.043523737 -0.52989066 -0.396595877  0.86015554
[42,]  0.27508460 -0.40186537  0.178730174  0.56727738  0.006432696 -0.39254923
[43,]  1.35873634  1.00322527  0.162637553  0.22572687 -0.307733044  0.15656029
[44,]  0.69705772 -1.40038695  1.278060619  0.18133862 -0.674257459 -0.44954973
[45,]  3.97496014  1.05380785  0.147161328  0.34944795 -0.252280487 -0.04917978
[46,] -0.43987724 -1.27115334 -0.282992648  1.19786748 -0.205306871  0.51846090
[47,] -0.39208658  0.73806851 -0.982719736 -0.15944357 -0.355743507 -0.39994439
[48,] -1.03423170 -0.99052635  0.256697930  0.36590892 -0.105225254 -0.83441955
[49,] -4.62244877  0.99651436 -0.235552453 -0.72406053 -0.289312211 -0.46508245
[50,] -1.20120463 -0.65135618 -0.651776626  0.50995398  0.238315484 -0.04788816
[51,] -2.01162212  1.42323676 -0.206399393  0.04870854  0.054556908  0.08238419
[52,] -1.95289970  0.75714295 -0.390827079 -0.09766184 -0.171795773  0.07208136
```

图 4：主成分得分

f)提取的主成分载荷矩阵如图 5 所示：

```
> loadings(pr)

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
x1  0.412  0.376  0.216  0.788         0.145
x2  0.381  0.357 -0.806 -0.118 -0.212 -0.141
x3  0.332  0.563  0.467 -0.588
x4 -0.461  0.279                -0.599  0.590
x5 -0.421  0.415 -0.250         0.738  0.205
x6 -0.430  0.407  0.146  0.134 -0.222 -0.749

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.167  0.167  0.167  0.167  0.167  0.167
Cumulative Var  0.167  0.333  0.500  0.667  0.833  1.000
```

图 5：主成分载荷矩阵

g)主成分散点图如图 6 所示：

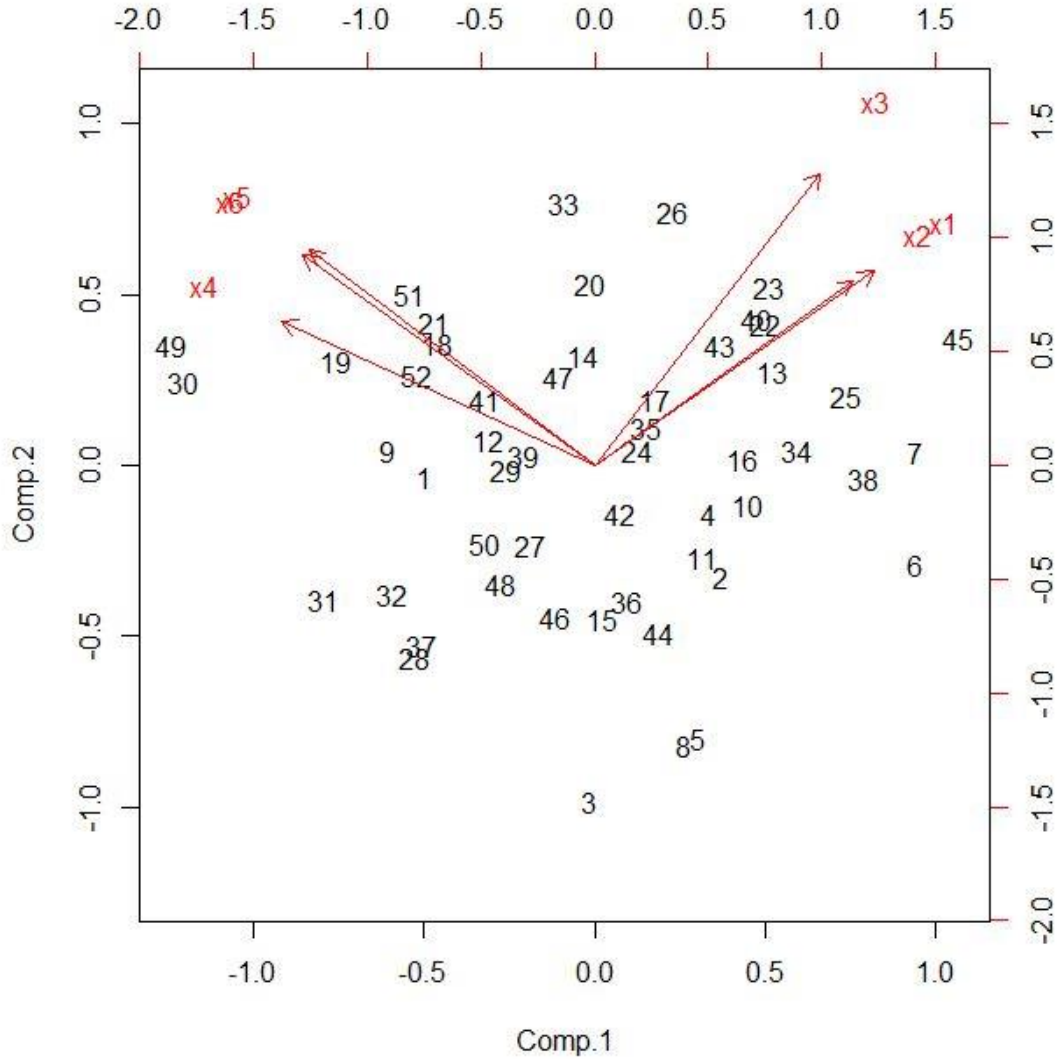


图 6：主成分散点图

答 2：

代码：

```
library(xlsx)
```

```
stusco<-read.xlsx("C:/Users/35088/Desktop/StuSco.xls",1)
(fa<-factanal(~.,factors=2,data=stusco,scores="regression"))
summary(fa)
```

结果如图 7 所示：

```
> library(xlsx)
> stusco<-read.xlsx("C:/Users/35088/Desktop/StuSco.xls",1)
> (fa<-factanal(~.,factors=2,data=stusco,scores="regression"))

Call:
factanal(x = ~., factors = 2, data = stusco, scores = "regression")

Uniquenesses:
   x1    x2    x3    x4    x5    x6
0.228 0.459 0.333 0.148 0.210 0.150

Loadings:
   Factor1 Factor2
x1 -0.309   0.823
x2 -0.309   0.668
x3          0.811
x4  0.848  -0.363
x5  0.862  -0.216
x6  0.899  -0.206

                Factor1 Factor2
SS loadings       2.471   2.001
Proportion Var    0.412   0.333
Cumulative Var    0.412   0.745

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 3.64 on 4 degrees of freedom.
The p-value is 0.457
```

图 7：因子分析结果

由图 7 知 Factor1 对变量$x_1$、$x_2$、$x_3$、$x_4$、$x_5$、$x_6$的总方差贡献为 2.471，Factor2 对变量$x_1$、$x_2$、$x_3$、$x_4$、$x_5$、$x_6$的总方差贡献为 2.001,Factor1 和 Factor2 的累积方差贡献率为 74.5%.

由于因子 Factor1 后几个变量的载荷因子接近于 1，这些变量涉及的科目是文科（语文、历史、英语），因此可称 Factor1 是文科因子；由于因子 Factor2 前几个变量的载荷因子接近于 1,这些变量涉及的科目是理科（数学、物理、化学），因此可称 Factor1 是理科因子。

5、城镇居民家庭人均消费性支出在衡量地区社会经济发展有着重要作用。现收集了 2011 年全国 31 个省、市、自治区反映城镇居民家庭人均消费性支出的八个主要指标数据，分别是食品、衣着、居住、家用设备用品、交通通信、文教娱乐、医疗保健和其他支出(X1-X8)，形成文件 StatAnn.xls。试利用这个数据集尝试 a)用谱系聚类法将 31 个省市区聚类，作谱系图。b)用 k 均值聚类方法将 31 个省市区聚成 4 个类别。

答 1：

代码：

```
library(xlsx)
df<-read.xlsx("C:/Users/35088/Desktop/StatAnn.xls",1)
df1<-scale(df[,-1],center=TRUE,scale=TRUE)
d<-dist(df1,method = "euclidean",diag = FALSE,upper = FALSE)
k1<-hclust(d,method = "ward.D2",members = NULL)#离差平方和法
plot(k1,labels=NULL,hang=0.1)
cophenetic(k1)
```

谱系图如图 8 所示：

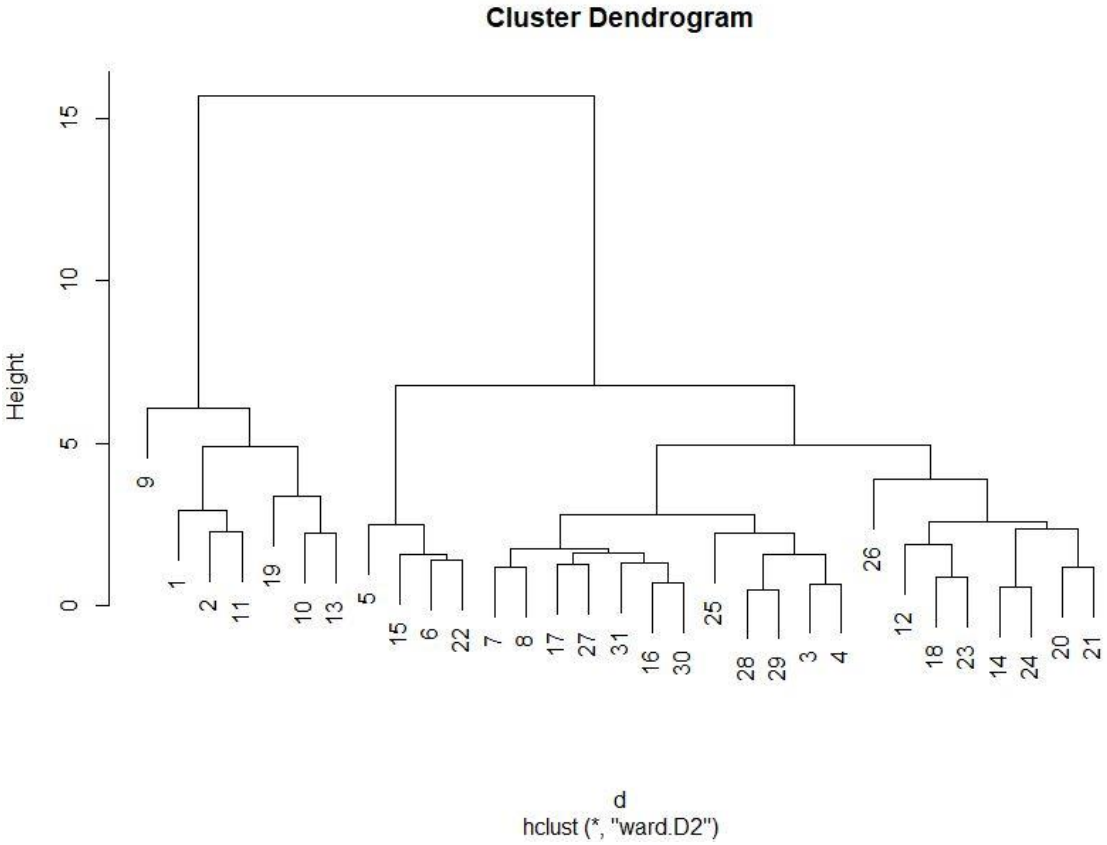


**图 8：聚类谱系图**

其中距离函数使用了欧式距离方法，聚类中使用了离差平方和法方法。

答 2：

代码：

```
library(xlsx)
df<-read.xlsx("C:/Users/35088/Desktop/StatAnn.xls",1)
(km<-kmeans(scale(df[,-1]),centers = 4))
```

K 均值聚类结果如图 9 所示：

```
> library(xlsx)
> df<-read.xlsx("C:/Users/35088/Desktop/StatAnn.xls",1)
> (km<-kmeans(scale(df[,-1]),centers = 4))
K-means clustering with 4 clusters of sizes 18, 8, 1, 4

Cluster means:
           x1         x2         x3         x4          x5         x6         x7         x8
1 -0.51340138 -0.5503531 -0.5321160 -0.5762527 -0.54860343 -0.5686669 -0.5513180 -0.5836120
2  0.03774279  0.7295235  0.1366747  0.2960442 -0.01206255  0.1593443  0.4191549  0.2053213
3  3.06829969  1.1730080  2.7361923  3.0752973  2.42881159  2.9725064  0.8204628  3.7113650
4  1.46774572  0.7242902  1.4371244  1.2322245  1.88563763  1.4971857  1.4375055  1.2877701

Clustering vector:
 [1] 4 4 1 1 2 2 2 1 3 2 4 1 2 1 2 1 1 1 4 1 1 2 1 1 1 1 2 1 1 1 1

Within cluster sum of squares by cluster:
[1] 36.12736 19.42740  0.00000 13.21290
 (between_SS / total_SS =  71.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

**图 9：k 均值聚类结果**

图 9 中的 Clustering vector 显示了 K 均值聚类结果，对应文件 StatAnn.xls，即可知道 31 个省市区所属类别。

6、数据文件 Cred.xls 是某金融机构一些个人客户的信用评估数据资料，现用此数据集建立客户信用度评价模型。Cred.xls 含有指标包括客户月收入、月生活支出、居住所有权(1:自有，0:租用)、目前工作的工作年限、前一工作的工作年限、目前住所的住所年限、前一住所的住所年限、家庭赡养人口数(X1-X8)，以及信用度评价(G:5 分制，最高 5)。试尝试用 Fisher 判别法建立客户信用度评价模型，并给出所建模型用于训练集 Cred.xls 所得的混淆矩阵。又，若有新客户其八项指标为(2500,1500,0,3,2,3,4,1)，试对该客户信用度进行评价。

答 1：

代码：

```
Cred<-read.xlsx("C:/Users/35088/Desktop/Cred.xls",1)
cred<-Cred[,-1]
library(MASS)
cred.lda<-lda(G~.,cred)
G.pre<-predict(cred.lda)
table(cred$G,G.pre$class) # 展示混淆矩阵
newdata=data.frame(x1=2500,x2=1500,x3=0,x4=3,x5=2,x6=3,x7=4,x8=1)
predict(cred.lda,newdata = newdata)$class
```

结果如图 10 所示：

```
> Cred<-read.xlsx("C:/Users/35088/Desktop/Cred.xls",1)
> cred<-Cred[,-1]
> library(MASS)
> cred.lda<-lda(G~.,cred)
> G.pre<-predict(cred.lda)
> table(cred$G,G.pre$class) # 展示混淆矩阵

   1 2 3 4 5
1  5 0 0 0 0
2  0 2 0 0 0
3  0 0 3 0 0
4  0 0 0 2 1
5  0 0 0 0 4
> newdata=data.frame(x1=2500,x2=1500,x3=0,x4=3,x5=2,x6=3,x7=4,x8=1)
> predict(cred.lda,newdata = newdata)$class
[1] 1
Levels: 1 2 3 4 5
```

图 10：判别分析结果

答 2：

使用 predict()函数对新用户的信用度进行评价，可知其信用度为 1.

## 五、实验总结

通过这次实验，我学习了主成分分析、聚类分析、判别分析的原理及在 R 中有关函数的用法，并通过实践操作，加深了理解和掌握。