# 第二章答案：

Ex2.1

x<-c(1,2,3)

y<-c(4,5,6)

e<-c(1,1,1)

z=2\*x+y+e

z1=crossprod(x,y)#z1为x1与x2的内积 或者 x%\*%y

z2=tcrossprod(x,y)#z1为x1与x2的外积 或者 x%o%y

z;z1;z2

要点：基本的列表赋值方法，内积和外积概念。内积为标量，外积为矩阵。

Ex2.2

A<-matrix(1:20,c(4,5));A

B<-matrix(1:20,nrow=4,byrow=TRUE);B

C=A+B;C

#不存在AB这种写法

E=A\*B;E

F<-A[1:3,1:3];F

H<-matrix(c(1,2,4,5),nrow=1);H

#H起过渡作用，不规则的数组下标

G<-B[,H];G

要点：矩阵赋值方法。默认是byrow=FALSE,数据按列放置。

取出部分数据的方法。可以用数组作为数组的下标取出数组元素。

Ex2.3

x<-c(rep(1,times=5),rep(2,times=3),rep(3,times=4),rep(4,times=2));x #或者省略times=，如下面的形式

x<-c(rep(1,5),rep(2,3),rep(3,4),rep(4,2));x

要点：rep（）的使用方法。rep（a,b）即将a重复b次

Ex2.4

n <- 5; H<-array(0,dim=c(n,n))

for (i in 1:n){for (j in 1:n){H[i,j]<-1/(i+j-1)}};H

G <- solve(H);G #求H的逆矩阵

ev <- eigen(H);ev #求H的特征值和特征向量

要点：数组初始化；for循环的使用

待解决：如何将很长的命令（如for循环）用几行打出来再执行？每次想换行的时候一按回车就执行了还没打完的命令...

Ex2.5

StudentData<-data.frame(name=c("zhangsan","lisi","wangwu","zhaoliu","dingyi"),sex=c("F","M","F","M","F"),age=c("14","15","16","14","15"),height=c("156","165","157","162","159"),weight=c("42","49","41.5","52","45.5"));StudentData

要点：数据框的使用

待解决：SSH登陆linux服务器中文显示乱码。此处用英文代替。

Ex2.6

write.table(StudentData,file="studentdata.txt")

#把数据框StudentData在工作目录里输出，输出的文件名为studentdata.txt.

StudentData\_a<-read.table("studentdata.txt");StudentData\_a

#以数据框的形式读取文档studentdata.txt，存入数据框StudentData\_a中。

write.csv(StudentData\_a,"studentdata.csv")

#把数据框StudentData\_a在工作目录里输出，输出的文件名为studentdata.csv,可用Excel打开.

要点：读写文件。read.table("file")

write.table(Rdata,"file")

read.csv("file")

write.csv(Rdata,"file")

外部文件，不论是待读入或是要写出的，命令中都得加双引号。

Ex2.7

Fun<-function(n){

if(n <= 0)

list(fail="please input a integer above 0!")

else{

repeat{

if(n==1) break

else if(n%%2==0){n<-n/2}

else n<- 3\*n+1

}

list("sucess!")

}

}

在linux下新建一个R文件，输入上述代码，保存为"2.7.R"

然后在当前目录下进入R环境，输入source("2.7.R")，即打开了这个程序脚本。

然后就可以执行函数了。

输入Fun(67)，显示

"sucess!"

输入Fun(-1)，显示

$fail

[1] "please input a integer above 0!"

待解决：source("\*.R")是可以理解为载入这个R文件吧？如何在R环境下关闭R文件呢？

OK,自己写的第一个R程序~~

# 3

Ex3.1

新建txt文件如下：3.1.txt

74.3 79.5 75.0 73.5 75.8 74.0 73.5 67.2 75.8 73.5 78.8 75.6 73.5 75.0 75.8

72.0 79.5 76.5 73.5 79.5 68.8 75.0 78.8 72.0 68.8 76.5 73.5 72.7 75.0 70.4

78.0 78.8 74.3 64.3 76.5 74.3 74.7 70.4 72.7 76.5 70.4 72.0 75.8 75.8 70.4

76.5 65.0 77.2 73.5 72.7 80.5 72.0 65.0 80.3 71.2 77.6 76.5 68.8 73.5 77.2

80.5 72.0 74.3 69.7 81.2 67.3 81.6 67.3 72.7 84.3 69.7 74.3 71.2 74.3 75.0

72.0 75.4 67.3 81.6 75.0 71.2 71.2 69.7 73.5 70.4 75.0 72.7 67.3 70.3 76.5

73.5 72.0 68.0 73.5 68.0 74.3 72.7 72.7 74.3 70.4

编写一个函数（程序名为data\_outline.R）描述样本的各种描述性统计量。

data\_outline<-function(x){

n<-length(x)

m<-mean(x)

v<-var(x)

s<-sd(x)

me<-median(x)

cv<-100\*s/m

css<-sum((x-m)^2)

uss<-sum(x^2)

R <- max(x)-min(x)

R1 <-quantile(x,3/4)-quantile(x,1/4)

sm <-s/sqrt(n)

g1 <-n/((n-1)\*(n-2))\*sum((x-m)^3)/s^3

g2 <-((n\*(n+1))/((n-1)\*(n-2)\*(n-3))\*sum((x-m)^4)/s^4-(3\*(n-1)^2)/((n-2)\*(n-3)))

data.frame(N=n,Mean=m,Var=v,std\_dev=s,Median=me,std\_mean=sm,CV=cv,CSS=css,USS=uss,R=R,R1=R1,Skewness=g1,Kurtosis=g2,row.names=1)

}

进入R，

source("data\_outline.R") #将程序调入内存

serumdata<-scan("3.1.txt");serumdata #将数据读入向量serumdata。

data\_outline(serumdata)

结果如下：

N Mean Var std\_dev Median std\_mean CV CSS USS R

1 100 73.696 15.41675 3.926417 73.5 0.3926417 5.327857 1526.258 544636.3 20

R1 Skewness Kurtosis

1 4.6 0.03854249 0.07051809

要点：read.table()用于读表格形式的文件。上述形式的数据由于第七行缺几个数据，故用read.table()不能读入。 scan()可以直接读纯文本文件。scan()和matrix()连用还可以将数据存放成矩阵形式。 X<-matrix(scan("3.1.txt",0),ncol=10,byrow=TRUE) #将上述数据放置成10\*10的矩阵。

scan()还可以从屏幕上直接输入数据。

Y<-scan()

然后按提示输入即可。结束输入时按回车即可。

Ex3.2

>hist(serumdata,freq=FALSE,col="purple",border="red",density=3,angle=60,main=paste("the histogram of serumdata"),xlab="age",ylab="frequency")#直方图。col是填充颜色。 默认空白。border是边框的颜色，默认前景色。density是在图上画条纹阴影，默认不画。angle是条纹阴影的倾斜角度（逆时针方向），默认45度。main, xlab, ylab是 标题，x和y坐标轴名称。

>lines(density(serumdata),col="blue")#密度估计曲线。

>x<-64:85

> lines(x,dnorm(x,mean(serumdata),sd(serumdata)),col="green") #正态分布的概率密度曲线

> plot(ecdf(serumdata),verticals=TRUE,do.p=FALSE) #绘制经验分布图

> lines(x,pnorm(x,mean(serumdata),sd(serumdata)),col="blue") #正态经验分布

> qqnorm(serumdata,col="purple") #绘制QQ图

> qqline(serumdata,col="red") #绘制QQ直线

Ex3.3

> stem(serumdata,scale=1) #作茎叶图。原始数据小数点后数值四舍五入。

The decimal point is at the |

64 | 300

66 | 23333

68 | 00888777

70 | 34444442222

72 | 0000000777777755555555555

74 | 033333333700000004688888

76 | 5555555226

78 | 0888555

80 | 355266

82 |

84 | 3

>boxplot(serumdata,col="lightblue",notch=T) #作箱线图。notch表示带有缺口。

> fivenum(serumdata) #五数总结

[1] 64.3 71.2 73.5 75.8 84.3

Ex3.4

> shapiro.test(serumdata) #正态性Shapori-Wilk检验方法

Shapiro-Wilk normality test

data: serumdata

W = 0.9897, p-value = 0.6437

结论：p值>0.05，可认为来自正态分布的总体。

> ks.test(serumdata,"pnorm",mean(serumdata),sd(serumdata)) #Kolmogrov-Smirnov检验，正态性

One-sample Kolmogorov-Smirnov test

data: serumdata

D = 0.0701, p-value = 0.7097

alternative hypothesis: two-sided

Warning message:

In ks.test(serumdata, "pnorm", mean(serumdata), sd(serumdata)) :

cannot compute correct p-values with ties

结论：p值>0.05，可认为来自正态分布的总体。

注意，这里的警告信息，是因为数据中有重复的数值，ks检验要求待检数据时连续的，不允许重复值。

Ex3.5

> y<-c(2,4,3,2,4,7,7,2,2,5,4,5,6,8,5,10,7,12,12,6,6,7,11,6,6,7,9,5,5,10,6,3,10) #输入数据

> f<-factor(c(rep(1,11),rep(2,10),rep(3,12))) #因子分类

> plot(f,y,col="lightgreen") #plot()生成箱线图

> x<-c(2,4,3,2,4,7,7,2,2,5,4)

> y<-c(5,6,8,5,10,7,12,12,6,6)

> z<-c(7,11,6,6,7,9,5,5,10,6,3,10)

> boxplot(x,y,z,names=c("1","2","3"),col=c(5,6,7)) #boxplot()生成箱线图

结论：第2和第3组没有显著差异。第1组合其他两组有显著差异。

Ex3.6

数据太多，懒得录入。离散图应该用plot即可。

Ex3.7

> studata<-read.table("3.7.txt") #读入数据

> data.frame(studata) #转化为数据框

V1 V2 V3 V4 V5 V6

1 1 alice f 13 56.5 84.0

2 2 becka f 13 65.3 98.0

3 3 gail f 14 64.3 90.0

4 4 karen f 12 56.3 77.0

5 5 kathy f 12 59.8 84.5

6 6 mary f 15 66.5 112.0

7 7 sandy f 11 51.3 50.5

8 8 sharon f 15 62.5 112.5

9 9 tammy f 14 62.8 102.5

10 10 alfred m 14 69.0 112.5

11 11 duke m 14 63.5 102.5

12 12 guido m 15 67.0 133.0

13 13 james m 12 57.3 83.0

14 14 jeffery m 13 62.5 84.0

15 15 john m 12 59.0 99.5

16 16 philip m 16 72.0 150.0

17 17 robert m 12 64.8 128.0

18 18 thomas m 11 57.5 85.0

19 19 william m 15 66.5 112.0

> names(studata)<-c("stuno","name","sex","age","height","weight"),studata #给各列命名

stuno name sex age height weight

1 1 alice f 13 56.5 84.0

2 2 becka f 13 65.3 98.0

3 3 gail f 14 64.3 90.0

...

> attach(studata) #将数据框调入内存

> plot(weight~height,col="red") #体重对于身高的散点图

> coplot(weight~height|sex,col="blue") #不同性别，体重与身高的散点图

> coplot(weight~height|age,col="blue") #不同年龄，体重与身高的散点图

> coplot(weight~height|age+sex,col="blue") #不同年龄和性别，体重与身高的散点图

Ex3.8

> x<-seq(-2,3,0.05)

> y<-seq(-1,7,0.05)

> f<-function(x,y) x^4-2\*x^2\*y+x^2-2\*x\*y+2\*y^2+4.5\*x-4\*y+4

> z<-outer(x,y,f) #必须做外积运算才能绘出三维图形

> contour(x,y,z,levels=c(0,1,2,3,4,5,10,15,20,30,40,50,60,80,100),col="blue") #二维等值线

> persp(x,y,z,theta=120,phi=0,expand=0.7,col="lightblue") #三位网格曲面

Ex3.9

> attach(studata)

> cor.test(height,weight) #Pearson相关性检验

Pearson's product-moment correlation

data: height and weight

t = 7.5549, df = 17, p-value = 7.887e-07

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.7044314 0.9523101

sample estimates:

cor

0.8777852

由此可见身高和体重是相关的。

Ex3.10

Ex3.11

上述两题原始数据太多，网上找不到，懒得录入。略。

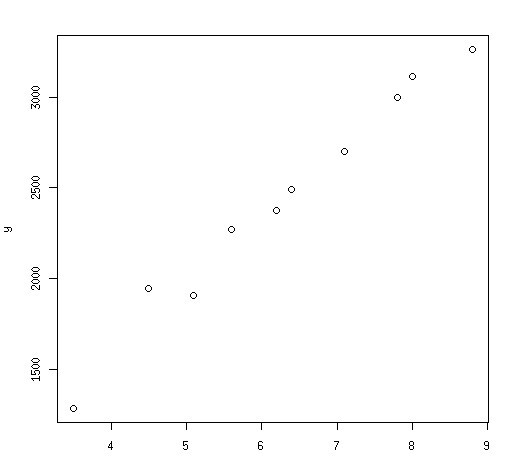
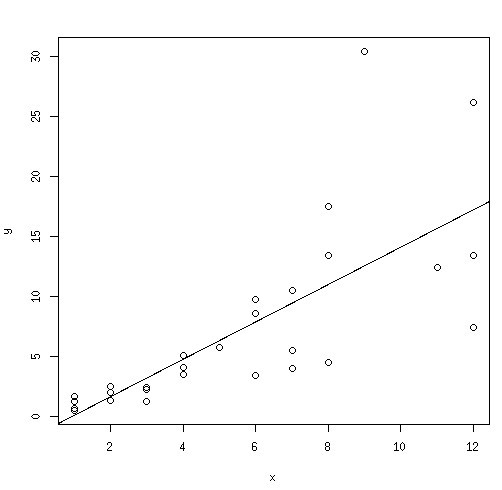
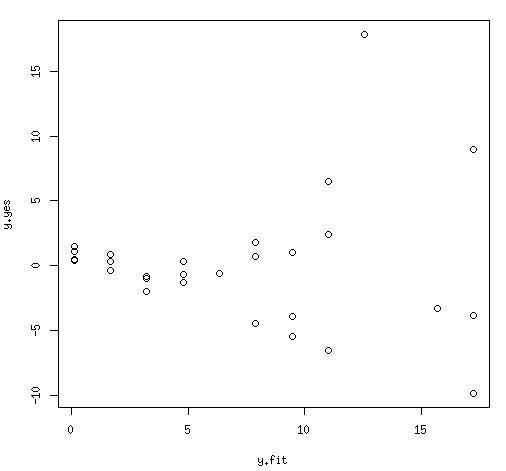
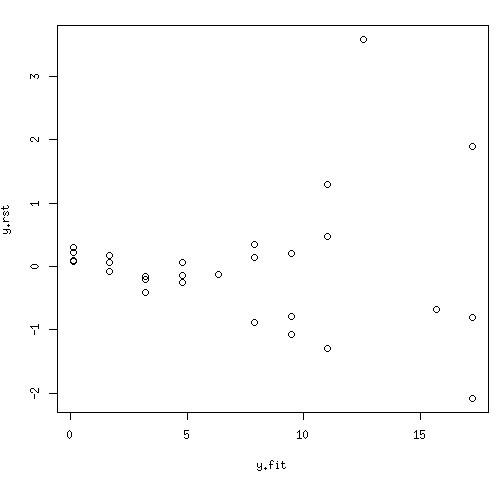
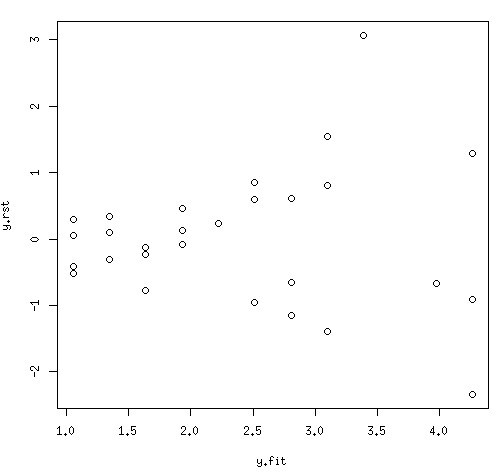
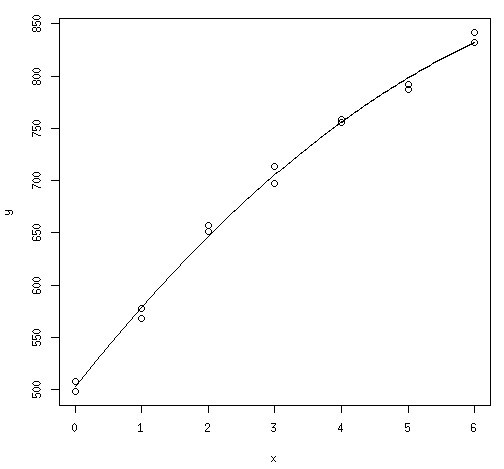
# 4

**Ex4.2**  
指数分布，λ的极大似然估计是n/sum(Xi)  
> x<-c(rep(5,365),rep(15,245),rep(25,150),rep(35,100),rep(45,70),rep(55,45),rep(65,25))  
> lamda<-length(x)/sum(x);lamda  
[1] 0.05  
  
**Ex4.3**  
Poisson分布P(x=k)=λ^k/k!\*e^(-λ)  
其均数和方差相等，均为λ，其含义为平均每升水中大肠杆菌个数。  
取均值即可。  
> x<-c(rep(0,17),rep(1,20),rep(2,10),rep(3,2),rep(4,1))  
> mean(x)  
[1] 1  
平均为1个。  
  
**Ex4.4**  
> obj<-function(x){f<-c(-13+x[1]+((5-x[2])\*x[2]-2)\*x[2],-29+x[1]+((x[2]+1)\*x[2]-14)\*x[2]) ;sum(f^2)}  #其实我也不知道这是在干什么。所谓的无约束优化问题。  
> x0<-c(0.5,-2)  
> **nlm**(obj,x0)  
$minimum  
[1] 48.98425  
  
$estimate  
[1] 11.4127791 -0.8968052  
  
$gradient  
[1]  1.411401e-08 -1.493206e-07  
  
$code  
[1] 1  
  
$iterations  
[1] 16  
  
**Ex4.5**  
> x<-c(54,67,68,78,70,66,67,70,65,69)  
> t.test(x)        #t.test()做单样本正态分布区间估计  
  
        One Sample t-test  
  
data:  x  
t = 35.947, df = 9, p-value = 4.938e-11  
alternative hypothesis: true mean is not equal to 0  
95 percent confidence interval:  
 63.1585 71.6415  
sample estimates:  
mean of x  
     67.4  
平均脉搏点估计为 67.4 ，95%区间估计为 63.1585 71.6415 。  
> t.test(x,alternative="less",mu=72)  #t.test()做单样本正态分布单侧区间估计  
  
        One Sample t-test  
data:  x  
t = -2.4534, df = 9, p-value = 0.01828  
alternative hypothesis: true mean is less than 72  
95 percent confidence interval:  
     -Inf 70.83705  
sample estimates:  
mean of x  
     67.4  
p值小于0.05，拒绝原假设，平均脉搏低于常人。  
要点：t.test()函数的用法。本例为单样本；可做双边和单侧检验。  
  
**Ex4.6**  
> x<-c(140,137,136,140,145,148,140,135,144,141);x  
 [1] 140 137 136 140 145 148 140 135 144 141  
> y<-c(135,118,115,140,128,131,130,115,131,125);y  
 [1] 135 118 115 140 128 131 130 115 131 125  
> t.test(x,y,var.equal=TRUE)  
  
        Two Sample t-test  
  
data:  x and y  
t = 4.6287, df = 18, p-value = 0.0002087  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
  7.53626 20.06374  
sample estimates:  
mean of x mean of y  
    140.6     126.8  
期望差的95%置信区间为  7.53626 20.06374 。  
要点：t.test()可做两正态样本均值差估计。此例认为两样本方差相等。  
ps：我怎么觉得这题应该用配对t检验？  
  
**Ex4.7**  
> x<-c(0.143,0.142,0.143,0.137)  
> y<-c(0.140,0.142,0.136,0.138,0.140)  
> t.test(x,y,var.equal=TRUE)  
  
        Two Sample t-test  
  
data:  x and y  
t = 1.198, df = 7, p-value = 0.2699  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -0.001996351  0.006096351  
sample estimates:  
mean of x mean of y  
  0.14125   0.13920  
 期望差的95%的区间估计为-0.001996351  0.006096351  
  
**Ex4.8**  
接Ex4.6  
> var.test(x,y)  
  
        F test to compare two variances  
  
data:  x and y  
F = 0.2353, num df = 9, denom df = 9, p-value = 0.04229  
alternative hypothesis: true ratio of variances is not equal to 1  
95 percent confidence interval:  
 0.05845276 0.94743902  
sample estimates:  
ratio of variances  
         0.2353305  
要点：var.test可做两样本方差比的估计。基于此结果可认为方差不等。  
因此，在Ex4.6中，计算期望差时应该采取方差不等的参数。  
> t.test(x,y)  
  
        Welch Two Sample t-test  
  
data:  x and y  
t = 4.6287, df = 13.014, p-value = 0.0004712  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
  7.359713 20.240287  
sample estimates:  
mean of x mean of y  
    140.6     126.8  
期望差的95%置信区间为 7.359713 20.240287 。  
要点：t.test(x,y，var.equal=TRUE)做方差相等的两正态样本的均值差估计  
      t.test(x,y)做方差不等的两正态样本的均值差估计  
  
**Ex4.9**  
> x<-c(rep(0,7),rep(1,10),rep(2,12),rep(3,8),rep(4,3),rep(5,2))  
> n<-length(x)  
> tmp<-sd(x)/sqrt(n)\*qnorm(1-0.05/2)  
> mean(x)  
[1] 1.904762  
> mean(x)-tmp;mean(x)+tmp  
[1] 1.494041  
[1] 2.315483  
平均呼唤次数为1.9  
0.95的置信区间为1.49,2,32  
  
  
**Ex4.10**  
> x<-c(1067,919,1196,785,1126,936,918,1156,920,948)  
> t.test(x,alternative="greater")  
  
        One Sample t-test  
  
data:  x  
t = 23.9693, df = 9, p-value = 9.148e-10  
alternative hypothesis: true mean is greater than 0  
95 percent confidence interval:  
 920.8443      Inf  
sample estimates:  
mean of x  
    997.1  
灯泡平均寿命置信度95%的单侧置信下限为 920.8443   
要点：t.test()做单侧置信区间估计

# 5

**Ex5.1**  
> x<-c(220, 188, 162, 230, 145, 160, 238, 188, 247, 113, 126, 245, 164, 231, 256, 183, 190, 158, 224, 175)  
> t.test(x,mu=225)  
  
        One Sample t-test  
  
data:  x  
t = -3.4783, df = 19, p-value = 0.002516  
alternative hypothesis: true mean is not equal to 225  
95 percent confidence interval:  
 172.3827 211.9173  
sample estimates:  
mean of x  
   192.15  
原假设：油漆工人的血小板计数与正常成年男子无差异。  
备择假设：油漆工人的血小板计数与正常成年男子有差异。  
p值小于0.05，拒绝原假设，认为油漆工人的血小板计数与正常成年男子有差异。  
  
上述检验是双边检验。也可采用单边检验。 备择假设：油漆工人的血小板计数小于正常成年男子。  
> t.test(x,mu=225,alternative="less")  
  
        One Sample t-test  
  
data:  x  
t = -3.4783, df = 19, p-value = 0.001258  
alternative hypothesis: true mean is less than 225  
95 percent confidence interval:  
     -Inf 208.4806  
sample estimates:  
mean of x  
   192.15  
同样可得出油漆工人的血小板计数小于正常成年男子的结论。  
  
**Ex5.2**  
> pnorm(1000,mean(x),sd(x))  
[1] 0.5087941  
> x  
 [1] 1067  919 1196  785 1126  936  918 1156  920  948  
> pnorm(1000,mean(x),sd(x))  
[1] 0.5087941  
x<=1000的概率为0.509,故x大于1000的概率为0.491.  
要点：pnorm计算正态分布的分布函数。在R软件中，计算值均为下分位点。  
  
**Ex5.3**  
> A<-c(113,120,138,120,100,118,138,123)  
> B<-c(138,116,125,136,110,132,130,110)  
> t.test(A,B,paired=TRUE)  
  
        Paired t-test  
  
data:  A and B  
t = -0.6513, df = 7, p-value = 0.5357  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -15.62889   8.87889  
sample estimates:  
mean of the differences  
                 -3.375  
p值大于0.05，接受原假设，两种方法治疗无差异。  
  
**Ex5.4**  
（1）  
**正态性W检验：**  
>x<-c(-0.7,-5.6,2,2.8,0.7,3.5,4,5.8,7.1,-0.5,2.5,-1.6,1.7,3,0.4,4.5,4.6,2.5,6,-1.4)  
>y<-c(3.7,6.5,5,5.2,0.8,0.2,0.6,3.4,6.6,-1.1,6,3.8,2,1.6,2,2.2,1.2,3.1,1.7,-2)                      
> shapiro.test(x)  
  
        Shapiro-Wilk normality test  
  
data:  x  
W = 0.9699, p-value = 0.7527  
  
> shapiro.test(y)  
  
        Shapiro-Wilk normality test  
  
data:  y  
W = 0.971, p-value = 0.7754  
**ks检验：**  
> ks.test(x,"pnorm",mean(x),sd(x))  
  
        One-sample Kolmogorov-Smirnov test  
  
data:  x  
D = 0.1065, p-value = 0.977  
alternative hypothesis: two-sided  
  
Warning message:  
In ks.test(x, "pnorm", mean(x), sd(x)) :  
  cannot compute correct p-values with ties  
> ks.test(y,"pnorm",mean(y),sd(y))  
  
        One-sample Kolmogorov-Smirnov test  
  
data:  y  
D = 0.1197, p-value = 0.9368  
alternative hypothesis: two-sided  
  
Warning message:  
In ks.test(y, "pnorm", mean(y), sd(y)) :  
  cannot compute correct p-values with ties  
**pearson拟合优度检验**，以x为例。  
> sort(x)  
 [1] -5.6 -1.6 -1.4 -0.7 -0.5  0.4  0.7  1.7  2.0  2.5  2.5  2.8  3.0  3.5  4.0  
[16]  4.5  4.6  5.8  6.0  7.1  
> x1<-table(cut(x,br=c(-6,-3,0,3,6,9)))  
> p<-pnorm(c(-3,0,3,6,9),mean(x),sd(x))  
> p  
[1] 0.04894712 0.24990009 0.62002288 0.90075856 0.98828138  
> p<-c(p[1],p[2]-p[1],p[3]-p[2],p[4]-p[3],1-p[4]);p  
[1] 0.04894712 0.20095298 0.37012278 0.28073568 0.09924144  
> chisq.test(x1,p=p)  
  
        Chi-squared test for given probabilities  
  
data:  x1  
X-squared = 0.5639, df = 4, p-value = 0.967  
  
Warning message:  
In chisq.test(x1, p = p) : Chi-squared approximation may be incorrect  
p值为0.967，接受原假设，x符合正态分布。  
  
（2）  
**方差相同模型t检验：**  
> t.test(x,y,var.equal=TRUE)  
  
        Two Sample t-test  
  
data:  x and y  
t = -0.6419, df = 38, p-value = 0.5248  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -2.326179  1.206179  
sample estimates:  
mean of x mean of y  
    2.065     2.625  
**方差不同模型t检验：**  
> t.test(x,y)  
  
        Welch Two Sample t-test  
  
data:  x and y  
t = -0.6419, df = 36.086, p-value = 0.525  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -2.32926  1.20926  
sample estimates:  
mean of x mean of y  
    2.065     2.625  
**配对t检验：**  
> t.test(x,y,paired=TRUE)  
  
        Paired t-test  
  
data:  x and y  
t = -0.6464, df = 19, p-value = 0.5257  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -2.373146  1.253146  
sample estimates:  
mean of the differences  
                  -0.56  
三种检验的结果都显示两组数据均值无差异。  
  
（3）  
**方差检验：**  
> var.test(x,y)  
  
        F test to compare two variances  
  
data:  x and y  
F = 1.5984, num df = 19, denom df = 19, p-value = 0.3153  
alternative hypothesis: true ratio of variances is not equal to 1  
95 percent confidence interval:  
 0.6326505 4.0381795  
sample estimates:  
ratio of variances  
          1.598361  
接受原假设，两组数据方差相同。  
  
**Ex5.5**  
> a <- c(126,125,136,128,123,138,142,116,110,108,115,140)  
> b <- c(162,172,177,170,175,152,157,159,160,162)  
**正态性检验，采用ks检验：**  
> ks.test(a,"pnorm",mean(a),sd(a))  
  
        One-sample Kolmogorov-Smirnov test  
  
data:  a  
D = 0.1464, p-value = 0.9266  
alternative hypothesis: two-sided  
  
> ks.test(b,"pnorm",mean(b),sd(b))  
  
        One-sample Kolmogorov-Smirnov test  
  
data:  b  
D = 0.2222, p-value = 0.707  
alternative hypothesis: two-sided  
  
Warning message:  
In ks.test(b, "pnorm", mean(b), sd(b)) :  
  cannot compute correct p-values with ties  
a和b都服从正态分布。  
**方差齐性检验：**  
> var.test(a,b)  
  
        F test to compare two variances  
  
data:  a and b  
F = 1.9646, num df = 11, denom df = 9, p-value = 0.3200  
alternative hypothesis: true ratio of variances is not equal to 1  
95 percent confidence interval:  
 0.5021943 7.0488630  
sample estimates:  
ratio of variances  
          1.964622  
可认为a和b的方差相同。  
**选用方差相同模型t检验：**  
> t.test(a,b,var.equal=TRUE)  
  
        Two Sample t-test  
  
data:  a and b  
t = -8.8148, df = 20, p-value = 2.524e-08  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -48.24975 -29.78358  
sample estimates:  
mean of x mean of y  
 125.5833  164.6000  
可认为两者有差别。  
  
**Ex5.6**  
二项分布总体的假设检验：  
> binom.test(57,400,p=0.147)  
  
        Exact binomial test  
  
data:  57 and 400  
number of successes = 57, number of trials = 400, p-value = 0.8876  
alternative hypothesis: true probability of success is not equal to 0.147  
95 percent confidence interval:  
 0.1097477 0.1806511  
sample estimates:  
probability of success  
                0.1425  
P 值>0.05，故接受原假设，表示调查结果支持该市老年人口的看法  
  
**Ex5.7**  
二项分布总体的假设检验：  
> binom.test(178,328,p=0.5,alternative="greater")  
  
        Exact binomial test  
  
data:  178 and 328  
number of successes = 178, number of trials = 328, p-value = 0.06794  
alternative hypothesis: true probability of success is greater than 0.5  
95 percent confidence interval:  
 0.4957616 1.0000000  
sample estimates:  
probability of success  
             0.5426829  
不能认为这种处理能增加母鸡的比例。  
  
**Ex5.8**  
利用pearson卡方检验是否符合特定分布：  
> **chisq.test(c(315,101,108,32),p=c(9,3,3,1)/16)**  
  
        Chi-squared test for given probabilities  
  
data:  c(315, 101, 108, 32)  
X-squared = 0.47, df = 3, p-value = 0.9254  
接受原假设，符合自由组合定律。  
  
**Ex5.9**  
利用pearson卡方检验是否符合泊松分布：  
> n<-length(z)  
> y<-c(92,68,28,11,1,0)  
> x<-0:5  
> q<-ppois(x,mean(rep(x,y)));n<-length(y)  
> p[1]<-q[1];p[n]=1-q[n-1]  
> **chisq.test(y,p=p)**  
  
        Chi-squared test for given probabilities  
  
data:  y  
X-squared = 2.1596, df = 5, p-value = 0.8267  
  
Warning message:  
In chisq.test(y, p = p) : Chi-squared approximation may be incorrect  
重新分组，合并频数小于5的组：  
> z<-c(92,68,28,12)  
> n<-length(z);p<-p[1:n-1];p[n]<-1-q[n-1]  
> chisq.test(z,p=p)  
  
        Chi-squared test for given probabilities  
  
data:  z  
X-squared = 0.9113, df = 3, p-value = 0.8227  
可认为数据服从泊松分布。  
  
**Ex5.10**  
ks检验 两个分布是否相同：  
> x<-c(2.36,3.14,752,3.48,2.76,5.43,6.54,7.41)  
> y<-c(4.38,4.25,6.53,3.28,7.21,6.55)  
> ks.test(x,y)  
  
        Two-sample Kolmogorov-Smirnov test  
  
data:  x and y  
D = 0.375, p-value = 0.6374  
alternative hypothesis: two-sided  
  
**Ex5.11**  
列联数据的独立性检验：  
> x <- c(358,2492,229,2745)  
> dim(x)<-c(2,2)  
> **chisq.test(x)**  
  
        Pearson's Chi-squared test with Yates' continuity correction  
  
data:  x  
X-squared = 37.4143, df = 1, p-value = 9.552e-10  
P 值<0.05 ，拒绝原假设，有影响。  
  
**Ex5.12**  
列联数据的独立性检验：  
> y  
     [,1] [,2] [,3]  
[1,]   45   12   10  
[2,]   46   20   28  
[3,]   28   23   30  
[4,]   11   12   35  
**> chisq.test(y)**  
  
        Pearson's Chi-squared test  
  
data:  y  
X-squared = 40.401, df = 6, p-value = 3.799e-07  
P 值<0.05 ，拒绝原假设，不独立，有关系。  
  
**Ex5.13**  
因有的格子的频数小于5，故采用fiser确切概率法检验独立性。  
> **fisher.test(x)**  
  
        Fisher's Exact Test for Count Data  
  
data:  x  
p-value = 0.6372  
alternative hypothesis: true odds ratio is not equal to 1  
95 percent confidence interval:  
 0.04624382 5.13272210  
sample estimates:  
odds ratio  
  0.521271  
p值大于0.05，两变量独立，两种工艺对产品的质量没有影响。  
  
**Ex5.14**  
由于是在相同个体上的两次试验，故采用McNemar检验。  
> **mcnemar.test(x)**  
  
        McNemar's Chi-squared test  
  
data:  x  
McNemar's chi-squared = 2.8561, df = 3, p-value = 0.4144  
p值大于0.05，不能认定两种方法测定结果不同。  
  
**Ex5.15**  
**符号检验**：  
H0：中位数>=14.6;  
H1: 中位数<14.6  
> x<-c(13.32,13.06,14.02,11.86,13.58,13.77,13.51,14.42,14.44,15.43)  
> **binom.test(sum(x)>14.6,length(x),al="l")**  
  
        Exact binomial test  
  
data:  sum(x) > 14.6 and length(x)  
number of successes = 1, number of trials = 10, p-value = 0.01074  
alternative hypothesis: true probability of success is less than 0.5  
95 percent confidence interval:  
 0.0000000 0.3941633  
sample estimates:  
probability of success  
                   0.1  
拒绝原假设，中位数小于14.6  
  
**Wilcoxon符号秩检验：**> **wilcox.test(x,mu=14.6,al="l",exact=F)**  
  
        Wilcoxon signed rank test with continuity correction  
  
data:  x  
V = 4.5, p-value = 0.01087  
alternative hypothesis: true location is less than 14.6  
拒绝原假设，中位数小于14.6  
  
**Ex5.16**  
**符号检验法：**  
> x<-c(48,33,37.5,48,42.5,40,42,36,11.3,22,36,27.3,14.2,32.1,52,38,17.3,20,21,46.1)  
> y<-c(37,41,23.4,17,31.5,40,31,36,5.7,11.5,21,6.1,26.5,21.3,44.5,28,22.6,20,11,22.3)  
> binom.test(sum(x>y),length(x))  
  
        Exact binomial test  
  
data:  sum(x > y) and length(x)  
number of successes = 14, number of trials = 20, p-value = 0.1153  
alternative hypothesis: true probability of success is not equal to 0.5  
95 percent confidence interval:  
 0.4572108 0.8810684  
sample estimates:  
probability of success  
                   0.7  
接受原假设，无差别。  
**Wilcoxon符号秩检验：**  
> wilcox.test(x,y,paired=TRUE,exact=FALSE)  
  
        Wilcoxon signed rank test with continuity correction  
  
data:  x and y  
V = 136, p-value = 0.005191  
alternative hypothesis: true location shift is not equal to 0  
拒绝原假设，有差别。  
**Wilcoxon秩和检验：**  
> wilcox.test(x,y,exact=FALSE)  
  
        Wilcoxon rank sum test with continuity correction  
  
data:  x and y  
W = 274.5, p-value = 0.04524  
alternative hypothesis: true location shift is not equal to 0  
拒绝原假设，有差别。  
**正态性检验：**  
> ks.test(x,"pnorm",mean(x),sd(x))  
  
        One-sample Kolmogorov-Smirnov test  
  
data:  x  
D = 0.1407, p-value = 0.8235  
alternative hypothesis: two-sided  
  
Warning message:  
In ks.test(x, "pnorm", mean(x), sd(x)) :  
  cannot compute correct p-values with ties  
> ks.test(y,"pnorm",mean(y),sd(y))  
  
        One-sample Kolmogorov-Smirnov test  
  
data:  y  
D = 0.1014, p-value = 0.973  
alternative hypothesis: two-sided  
两组数据均服从正态分布。  
**方差齐性检验：**  
> var.test(x,y)  
  
        F test to compare two variances  
  
data:  x and y  
F = 1.1406, num df = 19, denom df = 19, p-value = 0.7772  
alternative hypothesis: true ratio of variances is not equal to 1  
95 percent confidence interval:  
 0.4514788 2.8817689  
sample estimates:  
ratio of variances  
          1.140639  
可认为两组数据方差相同。  
综上，该数据可做t检验。  
**t检验：**  
> t.test(x,y,var.equal=TRUE)  
  
        Two Sample t-test  
  
data:  x and y  
t = 2.2428, df = 38, p-value = 0.03082  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
  0.812553 15.877447  
sample estimates:  
mean of x mean of y  
   33.215    24.870  
拒绝原假设，有差别。  
综上所述，Wilcoxon符号秩检验的差异检出能力最强，符号检验的差异检出最弱。  
  
**Ex5.17**  
**spearman秩相关检验：**  
> x<-c(24,17,20,41,52,23,46,18,15,20)  
> y<-c(8,1,4,7,9,5,10,3,2,6)  
> cor.test(x,y,method="spearman",exact=F)  
  
        Spearman's rank correlation rho  
  
data:  x and y  
S = 9.5282, p-value = 4.536e-05  
alternative hypothesis: true rho is not equal to 0  
sample estimates:  
      rho  
0.9422536  
**kendall秩相关检验：**  
> cor.test(x,y,method="kendall",exact=F)  
  
        Kendall's rank correlation tau  
  
data:  x and y  
z = 3.2329, p-value = 0.001225  
alternative hypothesis: true tau is not equal to 0  
sample estimates:  
      tau  
0.8090398  
二者有关系，呈正相关。  
  
**Ex5.18**  
> x<-rep(1:5,c(0,1,9,7,3));y<-rep(1:5,c(2,2,11,4,1))  
> wilcox.test(x,y,exact=F)  
  
        Wilcoxon rank sum test with continuity correction  
  
data:  x and y  
W = 266, p-value = 0.05509  
alternative hypothesis: true location shift is not equal to 0  
p值大于0.05，不能拒绝原假设，尚不能认为新方法的疗效显著优于原疗法。

# 6

**Ex6.1**  
（1）  
> x <- c(5.1, 3.5, 7.1, 6.2, 8.8, 7.8, 4.5, 5.6, 8.0, 6.4)  
> y <- c(1907, 1287, 2700, 2373, 3260, 3000, 1947, 2273, 3113,2493)  
> plot(x,y)  
[](http://photo.blog.sina.com.cn/showpic.html#blogid=681aaa5501012vkt&url=http://s14.sinaimg.cn/orignal/681aaa554ba006061415d)  
由此判断，Y和X有线性关系。  
（2）  
> lm.sol<-lm(y~1+x)  
> summary(lm.sol)  
  
Call:  
lm(formula = y ~ 1 + x)  
  
Residuals:  
     Min       1Q   Median       3Q      Max  
-128.591  -70.978   -3.727   49.263  167.228  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)     
(Intercept)   140.95     125.11   1.127    0.293     
x             364.18      19.26  18.908 6.33e-08 \*\*\*  
---  
Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 96.42 on 8 degrees of freedom  
Multiple R-squared: 0.9781,     Adjusted R-squared: 0.9754  
F-statistic: 357.5 on 1 and 8 DF,  p-value: 6.33e-08  
回归方程为 Y=140.95+364.18X  
（3）  
β1项很显著，但常数项β0不显著。  
回归方程很显著。  
（4）  
> new <- data.frame(x=7)  
> lm.pred<-predict(lm.sol,new,interval="prediction")  
> lm.pred  
       fit      lwr      upr  
1 2690.227 2454.971 2925.484  
故Y(7)= 2690.227, [2454.971,2925.484]  
  
**Ex6.2**  
(1)  
>pho<-data.frame(x1 <- c(0.4,0.4,3.1,0.6,4.7,1.7,9.4,10.1,11.6,12.6,10.9,23.1,23.1,21.6,23.1,1.9,26.8,29.9), x2 <- c(52,34,19,34,24,65,44,31,29,58,37,46,50,44,56,36,58,51), x3 <- c(158,163,37,157,59,123,46,117,173,112,111,114,134,73,168,143,202,124), y <- c(64,60,71,61,54,77,81,93,93,51,76,96,77,93,95,54,168,99))  
> lm.sol<-lm(y~x1+x2+x3,data=pho)  
> summary(lm.sol)  
  
Call:  
lm(formula = y ~ x1 + x2 + x3, data = pho)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-27.575 -11.160  -2.799  11.574  48.808  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  44.9290    18.3408   2.450  0.02806 \*  
x1            1.8033     0.5290   3.409  0.00424 \*\*  
x2           -0.1337     0.4440  -0.301  0.76771    
x3            0.1668     0.1141   1.462  0.16573    
---  
Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 19.93 on 14 degrees of freedom  
Multiple R-squared: 0.551,      Adjusted R-squared: 0.4547  
F-statistic: 5.726 on 3 and 14 DF,  p-value: 0.009004  
回归方程为 y=44.9290+1.8033x1-0.1337x2+0.1668x3  
（2）  
回归方程显著，但有些回归系数不显著。  
（3）  
**> lm.step<-step(lm.sol)**  
Start:  AIC=111.2  
y ~ x1 + x2 + x3  
  
       Df Sum of Sq     RSS     AIC  
- x2    1      36.0  5599.4   109.3  
<none>               5563.4   111.2  
- x3    1     849.8  6413.1   111.8  
- x1    1    4617.8 10181.2   120.1  
  
Step:  AIC=109.32  
y ~ x1 + x3  
  
       Df Sum of Sq     RSS     AIC  
<none>               5599.4   109.3  
- x3    1     833.2  6432.6   109.8  
- x1    1    5169.5 10768.9   119.1  
  
**> summary(lm.step)**  
  
Call:  
lm(formula = y ~ x1 + x3, data = pho)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-29.713 -11.324  -2.953  11.286  48.679  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  41.4794    13.8834   2.988  0.00920 \*\*  
x1            1.7374     0.4669   3.721  0.00205 \*\*  
x3            0.1548     0.1036   1.494  0.15592    
---  
Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 19.32 on 15 degrees of freedom  
Multiple R-squared: 0.5481,     Adjusted R-squared: 0.4878  
F-statistic: 9.095 on 2 and 15 DF,  p-value: 0.002589  
x3仍不够显著。  
再用drop1函数做逐步回归。  
**> drop1(lm.step)**  
Single term deletions  
  
Model:  
y ~ x1 + x3  
       Df Sum of Sq     RSS     AIC  
<none>               5599.4   109.3  
x1      1    5169.5 10768.9   119.1  
x3      1     833.2  6432.6   109.8  
可以考虑再去掉x3.  
**> lm.opt<-lm(y~x1,data=pho);summary(lm.opt)**  
  
Call:  
lm(formula = y ~ x1, data = pho)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-31.486  -8.282  -1.674   5.623  59.337  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)     
(Intercept)  59.2590     7.4200   7.986 5.67e-07 \*\*\*  
x1            1.8434     0.4789   3.849  0.00142 \*\*  
---  
Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 20.05 on 16 degrees of freedom  
Multiple R-squared: 0.4808,     Adjusted R-squared: 0.4484  
F-statistic: 14.82 on 1 and 16 DF,  p-value: 0.001417  
皆显著。  
  
**Ex6.3**  
> x<-c(1,1,1,1,2,2,2,3,3,3,4,4,4,5,6,6,6,7,7,7,8,8,8,9,11,12,12,12)  
> y<-c(0.6,1.6,0.5,1.2,2.0,1.3,2.5,2.2,2.4,1.2,3.5,4.1,5.1,5.7,3.4,9.7,8.6,4.0,5.5,10.5,17.5,13.4,4.5,30.4,12.4,13.4,26.2,7.4)  
> plot(x,y)  
> lm.sol<-lm(y~1+x)  
> summary(lm.sol)  
  
Call:  
lm(formula = y ~ 1 + x)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-9.8413 -2.3369 -0.0214  1.0592 17.8320  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)     
(Intercept)  -1.4519     1.8353  -0.791    0.436     
x             1.5578     0.2807   5.549 7.93e-06 \*\*\*  
---  
Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 5.168 on 26 degrees of freedom  
Multiple R-squared: 0.5422,     Adjusted R-squared: 0.5246  
F-statistic:  30.8 on 1 and 26 DF,  p-value: 7.931e-06  
线性回归方程为 y=-1.4519+1.5578x，通过F 检验。 常数项参数未通过t 检验。  
> abline(lm.sol)  
[](http://photo.blog.sina.com.cn/showpic.html#blogid=681aaa5501012vkt&url=http://s3.sinaimg.cn/orignal/681aaa554ba29a2d63852)  
> y.yes<-resid(lm.sol)  
> y.fit<-predict(lm.sol)  
> y.rst<-rstandard(lm.sol)  
> plot(y.yes~y.fit)  
[](http://photo.blog.sina.com.cn/showpic.html#blogid=681aaa5501012vkt&url=http://s8.sinaimg.cn/orignal/681aaa554ba29c3200bf7)  
  
> plot(y.rst~y.fit)  
[](http://photo.blog.sina.com.cn/showpic.html#blogid=681aaa5501012vkt&url=http://s10.sinaimg.cn/orignal/681aaa554ba29c20b25b9)  
  
残差并非是等方差的。  
修正模型，对相应变量Y做开方。  
> lm.new<-update(lm.sol,sqrt(.)~.)  
> summary(lm.new)  
  
Call:  
lm(formula = sqrt(y) ~ x)  
  
Residuals:  
     Min       1Q   Median       3Q      Max  
-1.54255 -0.45280 -0.01177  0.34925  2.12486  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)     
(Intercept)  0.76650    0.25592   2.995  0.00596 \*\*  
x            0.29136    0.03914   7.444 6.64e-08 \*\*\*  
---  
Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.7206 on 26 degrees of freedom  
Multiple R-squared: 0.6806,     Adjusted R-squared: 0.6684  
F-statistic: 55.41 on 1 and 26 DF,  p-value: 6.645e-08  
此时所有参数和方程均通过检验。  
对新模型做标准化残差图，情况有所改善，不过还是存在一个离群值。第24和第28个值存在问题。  
[](http://photo.blog.sina.com.cn/showpic.html#blogid=681aaa5501012vkt&url=http://s11.sinaimg.cn/orignal/681aaa554ba29e4bc2f9a)  
  
**Ex6.4**  
> toothpaste<-data.frame( X1=c(-0.05, 0.25,0.60,0,0.20, 0.15,-0.15, 0.15,0.10,0.40,0.45,0.35,0.30, 0.50,0.50,0.40,-0.05,-0.05,-0.10,0.20,0.10,0.50,0.60,-0.05,0,0.05,0.55),X2=c(5.50,6.75,7.25,5.50,6.50,6.75,5.25,6.00,6.25,7.00,6.90,6.80,6.80,7.10,7.00,6.80,6.50,6.25,6.00,6.50,7.00,6.80,6.80,6.50,5.75,5.80,6.80),Y=c(7.38,8.51,9.52,7.50,8.28,8.75,7.10,8.00,8.15,9.10,8.86,8.90,8.87,9.26,9.00,8.75,7.95, 7.65,7.27,8.00,8.50,8.75,9.21,8.27,7.67,7.93,9.26))  
> lm.sol<-lm(Y~X1+X2,data=toothpaste); summary(lm.sol)  
  
Call:  
lm(formula = Y ~ X1 + X2, data = toothpaste)  
  
Residuals:  
     Min       1Q   Median       3Q      Max  
-0.37130 -0.10114  0.03066  0.10016  0.30162  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)     
(Intercept)   4.0759     0.6267   6.504 1.00e-06 \*\*\*  
X1            1.5276     0.2354   6.489 1.04e-06 \*\*\*  
X2            0.6138     0.1027   5.974 3.63e-06 \*\*\*  
---  
Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.1767 on 24 degrees of freedom  
Multiple R-squared: 0.9378,     Adjusted R-squared: 0.9327  
F-statistic:   181 on 2 and 24 DF,  p-value: 3.33e-15  
回归诊断：  
> influence.measures(lm.sol)  
Influence measures of  
         lm(formula = Y ~ X1 + X2, data = toothpaste) :  
  
     dfb.1\_   dfb.X1   dfb.X2   dffit cov.r   cook.d    hat inf  
1   0.00908  0.00260 -0.00847  0.0121 1.366 5.11e-05 0.1681     
2   0.06277  0.04467 -0.06785 -0.1244 1.159 5.32e-03 0.0537     
3  -0.02809  0.07724  0.02540  0.1858 1.283 1.19e-02 0.1386     
4   0.11688  0.05055 -0.11078  0.1404 1.377 6.83e-03 0.1843   \*  
5   0.01167  0.01887 -0.01766 -0.1037 1.141 3.69e-03 0.0384     
6  -0.43010 -0.42881  0.45774  0.6061 0.814 1.11e-01 0.0936     
7   0.07840  0.01534 -0.07284  0.1082 1.481 4.07e-03 0.2364   \*  
8   0.01577  0.00913 -0.01485  0.0208 1.237 1.50e-04 0.0823     
9   0.01127 -0.02714 -0.00364  0.1071 1.156 3.95e-03 0.0466     
10 -0.07830  0.00171  0.08052  0.1890 1.155 1.22e-02 0.0726     
11  0.00301 -0.09652 -0.00365 -0.2281 1.127 1.76e-02 0.0735     
12 -0.03114  0.01848  0.03459  0.1542 1.132 8.12e-03 0.0514     
13 -0.09236 -0.03801  0.09940  0.2201 1.071 1.62e-02 0.0522     
14 -0.02650  0.03434  0.02606  0.1179 1.235 4.81e-03 0.0956     
15  0.00968 -0.11445 -0.00857 -0.2545 1.150 2.19e-02 0.0910     
16 -0.00285 -0.06185  0.00098 -0.1608 1.146 8.83e-03 0.0594     
17  0.07201  0.09744 -0.07796 -0.1099 1.364 4.19e-03 0.1731     
18  0.15132  0.30204 -0.17755 -0.3907 1.087 5.04e-02 0.1085     
19  0.07489  0.47472 -0.12980 -0.7579 0.731 1.66e-01 0.1092     
20  0.05249  0.08484 -0.07940 -0.4660 0.625 6.11e-02 0.0384   \*  
21  0.07557  0.07284 -0.07861 -0.0880 1.471 2.69e-03 0.2304   \*  
22 -0.17959 -0.39016  0.18241 -0.5494 0.912 9.41e-02 0.1022     
23  0.06026  0.10607 -0.06207  0.1251 1.374 5.42e-03 0.1804     
24 -0.54830 -0.74197  0.59358  0.8371 0.914 2.13e-01 0.1731     
25  0.08541  0.01624 -0.07775  0.1314 1.249 5.97e-03 0.1069     
26  0.32556  0.11734 -0.30200  0.4480 1.018 6.49e-02 0.1033     
27  0.17243  0.32754 -0.17676  0.4127 1.148 5.66e-02 0.1369     
> source("Reg\_Diag.R");Reg\_Diag(lm.sol) #薛毅老师自己写的程序  
      residual s1    standard s2     student s3 hat\_matrix s4      DFFITS s5  
1   0.00443843     0.02753865     0.02695925    0.16811819     0.01211949    
2  -0.09114255    -0.53021138    -0.52211469    0.05369239    -0.12436727    
3   0.07726887     0.47112863     0.46335666    0.13857353     0.18584310    
4   0.04805665     0.30111062     0.29532912    0.18427663     0.14036860    
5  -0.09130271    -0.52689847    -0.51881406    0.03838430    -0.10365442    
6   0.30162101     1.79287913     1.88596579    0.09362223     0.60613406    
7   0.03066005     0.19855842     0.19453763    0.23641540  \*  0.10824626    
8   0.01199519     0.07085860     0.06937393    0.08226537     0.02077047    
9   0.08491891     0.49217591     0.48426323    0.04664158     0.10711246    
10  0.11625405     0.68315814     0.67537315    0.07261134     0.18897969    
11 -0.13874451    -0.81570765    -0.80983786    0.07348894    -0.22807820    
12  0.11540228     0.67051940     0.66263761    0.05137589     0.15420864    
13  0.16178406     0.94041623     0.93806144    0.05219432     0.22013204    
14  0.06210727     0.36957277     0.36282531    0.09557411     0.11794546    
15 -0.13650951    -0.81026658    -0.80428349    0.09101221    -0.25449541    
16 -0.11097950    -0.64757782    -0.63955524    0.05943308    -0.16076716    
17 -0.03939381    -0.24515626    -0.24029557    0.17309048    -0.10993940    
18 -0.18593575    -1.11438446    -1.12029013    0.10845395    -0.39073410    
19 -0.33609591    -2.01522068  \* -2.16439284  \* 0.10922236    -0.75789180  \*  
20 -0.37130271  \* -2.14274943  \* -2.33258738  \* 0.03838430    -0.46603012    
21 -0.02545527    -0.16420856    -0.16084153    0.23042354  \* -0.08801075    
22 -0.26374306    -1.57517595    -1.62848498    0.10217431    -0.54936198    
23  0.04349338     0.27187605     0.26656251    0.18041800     0.12506702    
24  0.28060619     1.74627363     1.82969510    0.17309048     0.83711731  \*  
25  0.06459859     0.38683016     0.37987153    0.10691352     0.13143357    
26  0.21752520     1.29995371     1.31989945    0.10329116     0.44796770    
27  0.16987516     1.03474390     1.03633781    0.13685835     0.41266341    
   cooks\_distance s6  COVRATIO s7  
1    5.108777e-05    1.3656752    
2    5.316885e-03    1.1589547    
3    1.190200e-02    1.2827036    
4    6.827446e-03    1.3771332    
5    3.693897e-03    1.1410104    
6    1.106753e-01    0.8143179    
7    4.068871e-03    1.4806452  \*  
8    1.500251e-04    1.2372586    
9    3.950358e-03    1.1560508    
10   1.218047e-02    1.1550557    
11   1.759216e-02    1.1271148    
12   8.116460e-03    1.1316638    
13   1.623390e-02    1.0710597    
14   4.811117e-03    1.2349272    
15   2.191171e-02    1.1501502    
16   8.832858e-03    1.1457602    
17   4.193532e-03    1.3637206    
18   5.035591e-02    1.0866343    
19   1.659840e-01    0.7313914    
20   6.109050e-02    0.6248838    
21   2.691197e-03    1.4714103    
22   9.412101e-02    0.9121786    
23   5.423856e-03    1.3735324    
24   2.127740e-01  \* 0.9139942    
25   5.971157e-03    1.2485557    
26   6.488529e-02    1.0178195    
27   5.658922e-02    1.1479080    
为什么两种方法检测结果不一样呢...不继续了  
  
**Ex6.5**  
> cement<-data.frame(X1=c( 7,  1, 11, 11,  7, 11,  3,  1,  2, 21,  1, 11, 10),X2=c(26, 29, 56, 31, 52, 55, 71, 31, 54, 47, 40, 66, 68),X3=c( 6, 15,  8,  8,  6,  9, 17, 22, 18,  4, 23,  9, 8),X4=c(60, 52, 20, 47, 33, 22,  6, 44, 22, 26, 34, 12, 12),Y =c(78.5, 74.3, 104.3,  87.6,  95.9, 109.2, 102.7, 72.5, 93.1,115.9,  83.8, 113.3, 109.4))  
> xx<-cor(cement[1:4])  
> kappa(xx,exact=T)  
[1] 1376.881  
大于1000，认为有严重的多重共线性。  
> eigen(xx)  
$values  
[1] 2.235704035 1.576066070 0.186606149 0.001623746  
  
$vectors  
           [,1]       [,2]       [,3]      [,4]  
[1,] -0.4759552  0.5089794  0.6755002 0.2410522  
[2,] -0.5638702 -0.4139315 -0.3144204 0.6417561  
[3,]  0.3940665 -0.6049691  0.6376911 0.2684661  
[4,]  0.5479312  0.4512351 -0.1954210 0.6767340  
即2.235704035X1+1.576066070X2+0.186606149X3+0.001623746X4=0  
可以忽略X4项，可以看出X1，X2，X3存在共线性。  
删去X3和X4后：  
> xx<-cor(cement[1:2])  
> kappa(xx,exact=T)  
[1] 1.59262  
共线性消失了。  
如果去掉X3呢？  
> cement1<-cbind(cement$X1,cement$X2,cement$X4)  
> xx<-cor(cement1)  
> kappa(xx,exact=T)  
[1] 77.25113  
如果去掉X4呢？  
> xx<-cor(cement[1:3])  
> kappa(xx,exact=T)  
[1] 11.12112  
看起来去掉X3和X4是合理的。  
  
**Ex6.6**  
>infection<-data.frame(x1<-c(0,1,0,0,0,1,1,1),x2<-c(0,0,1,0,1,0,1,1),x3<-c(0,0,0,1,1,1,0,1),success<-c(0,0,23,8,28,0,11,1),fail<-c(9,0,3,32,30,2,87,17))  
> infection$Ymat<-cbind(infection$success,infection$fail)  
> glm.sol<-glm(Ymat~x1+x2+x3,family=binomial,data=infection)  
> summary(glm.sol)  
  
Call:  
glm(formula = Ymat ~ x1 + x2 + x3, family = binomial, data = infection)  
  
Deviance Residuals:  
       1         2         3         4         5         6         7         8   
-2.56229   0.00000   1.49623   1.21563  -0.78520  -0.15231  -0.07162   0.26470   
  
Coefficients:  
            Estimate Std. Error z value Pr(>|z|)     
(Intercept)  -0.8207     0.4947  -1.659   0.0971 .   
x1           -3.2544     0.4813  -6.761 1.37e-11 \*\*\*  
x2            2.0299     0.4553   4.459 8.25e-06 \*\*\*  
x3           -1.0720     0.4254  -2.520   0.0117 \*   
---  
Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
    Null deviance: 83.491  on 6  degrees of freedom  
Residual deviance: 10.997  on 3  degrees of freedom  
AIC: 36.178  
  
Number of Fisher Scoring iterations: 5  
回归模型为  
P=exp(-0.8207-3.2544x1+2.0299x2-1.0720x3)/(1+exp(-0.8207-3.2544x1+2.0299x2-1.0720x3))  
  
**Ex6.7**  
（1）  
> x<-c(rep(0:6,rep(2,7)))  
> y<-c(508.1,498.4,568.2,577.3,651.7,657.0,713.4,697.5,755.3,758.9,787.6,792.1,841.4,831.8)  
> lm.sol<-lm(y~1+x)  
> summary(lm.sol)  
  
Call:  
lm(formula = y ~ 1 + x)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-25.400 -11.484  -3.779  14.629  24.921  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)     
(Intercept)  523.800      8.474   61.81  < 2e-16 \*\*\*  
x             54.893      2.350   23.36 2.26e-11 \*\*\*  
---  
Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 17.59 on 12 degrees of freedom  
Multiple R-squared: 0.9785,     Adjusted R-squared: 0.9767  
F-statistic: 545.5 on 1 and 12 DF,  p-value: 2.265e-11  
线性回归模型为y=523.800+54.893x,通过t检验和F检验。  
（2）  
> lm.sol<-lm(y~1+x+I(x^2));summary(lm.sol)  
  
Call:  
lm(formula = y ~ 1 + x + I(x^2))  
  
Residuals:  
     Min       1Q   Median       3Q      Max  
-10.6643  -5.6622  -0.4655   5.5000  10.6679  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)     
(Intercept) 502.5560     4.8500 103.619  < 2e-16 \*\*\*  
x            80.3857     3.7861  21.232 2.81e-10 \*\*\*  
I(x^2)       -4.2488     0.6063  -7.008 2.25e-05 \*\*\*  
---  
Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 7.858 on 11 degrees of freedom  
Multiple R-squared: 0.9961,     Adjusted R-squared: 0.9953  
F-statistic:  1391 on 2 and 11 DF,  p-value: 5.948e-14  
多项式回归模型为：  
y=502.5560+80.3857x-4.2488x^2,通过t检验和F检验。  
(3)作散点图和拟合曲线：  
> plot(x,y)  
> xfit<-seq(0,6,0.01)  
> yfit<-predict(lm.sol,data.frame(x=xfit))  
> lines(xfit,yfit)  
[](http://photo.blog.sina.com.cn/showpic.html#blogid=681aaa5501012vkt&url=http://s2.sinaimg.cn/orignal/681aaa554ba3f189c3811)  
  
**Ex6.8**  
读入数据：  
> cancer<-read.table("data",header=T)  
> cancer  
   x1 x2 x3 x4 x5 y  
1  70 64  5  1  1 1  
2  60 63  9  1  1 0  
3  70 65 11  1  1 0  
4  40 69 10  1  1 0  
5  40 63 58  1  1 0  
6  70 48  9  1  1 0  
7  70 48 11  1  1 0  
8  80 63  4  2  1 0  
9  60 63 14  2  1 0  
10 30 53  4  2  1 0  
11 80 43 12  2  1 0  
12 40 55  2  2  1 0  
13 60 66 25  2  1 1  
14 40 67 23  2  1 0  
15 20 61 19  3  1 0  
16 50 63  4  3  1 0  
17 50 66 16  0  1 0  
18 40 68 12  0  1 0  
19 80 41 12  0  1 1  
20 70 53  8  0  1 1  
21 60 37 13  1  1 0  
22 90 54 12  1  0 1  
23 50 52  8  1  0 1  
24 70 50  7  1  0 1  
25 20 65 21  1  0 0  
26 80 52 28  1  0 1  
27 60 70 13  1  0 0  
28 50 40 13  1  0 0  
29 70 36 22  2  0 0  
30 40 44 36  2  0 0  
31 30 54  9  2  0 0  
32 30 59 87  2  0 0  
33 40 69  5  3  0 0  
34 60 50 22  3  0 0  
35 80 62  4  3  0 0  
36 70 68 15  0  0 0  
37 30 39  4  0  0 0  
38 60 49 11  0  0 0  
39 80 64 10  0  0 1  
40 70 67 18  0  0 1  
> glm.sol<-glm(y~x1+x2+x3+x4+x5,family=binomial,data=cancer);summary(glm.sol)  
  
Call:  
glm(formula = y ~ x1 + x2 + x3 + x4 + x5, family = binomial,  
    data = cancer)  
  
Deviance Residuals:  
     Min        1Q    Median        3Q       Max   
-1.71500  -0.66725  -0.22254   0.09936   2.23936   
  
Coefficients:  
            Estimate Std. Error z value Pr(>|z|)   
(Intercept) -7.01140    4.47534  -1.567   0.1172   
x1           0.09994    0.04304   2.322   0.0202 \*  
x2           0.01415    0.04697   0.301   0.7631   
x3           0.01749    0.05458   0.320   0.7486   
x4          -1.08297    0.58721  -1.844   0.0651 .  
x5          -0.61309    0.96066  -0.638   0.5233   
---  
Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
    Null deviance: 44.987  on 39  degrees of freedom  
Residual deviance: 28.392  on 34  degrees of freedom  
AIC: 40.392  
  
Number of Fisher Scoring iterations: 6  
有的系数并不显著。  
下面做逐步回归：  
> glm.new<-step(glm.sol)  
Start:  AIC=40.39  
y ~ x1 + x2 + x3 + x4 + x5  
  
       Df Deviance    AIC  
- x3    1   28.484 38.484  
- x2    1   28.484 38.484  
- x5    1   28.799 38.799  
<none>      28.392 40.392  
- x4    1   32.642 42.642  
- x1    1   38.306 48.306  
  
Step:  AIC=38.48  
y ~ x1 + x2 + x4 + x5  
  
       Df Deviance    AIC  
- x2    1   28.564 36.564  
- x5    1   28.993 36.993  
<none>      28.484 38.484  
- x4    1   32.705 40.705  
- x1    1   38.478 46.478  
  
Step:  AIC=36.56  
y ~ x1 + x4 + x5  
  
       Df Deviance    AIC  
- x5    1   29.073 35.073  
<none>      28.564 36.564  
- x4    1   32.892 38.892  
- x1    1   38.478 44.478  
  
Step:  AIC=35.07  
y ~ x1 + x4  
  
       Df Deviance    AIC  
<none>      29.073 35.073  
- x4    1   33.535 37.535  
- x1    1   39.131 43.131  
只留下x1和x4两个变量。  
> summary(glm.new)  
  
Call:  
glm(formula = y ~ x1 + x4, family = binomial, data = cancer)  
  
Deviance Residuals:  
    Min       1Q   Median       3Q      Max   
-1.4825  -0.6617  -0.1877   0.1227   2.2844   
  
Coefficients:  
            Estimate Std. Error z value Pr(>|z|)   
(Intercept) -6.13755    2.73844  -2.241   0.0250 \*  
x1           0.09759    0.04079   2.393   0.0167 \*  
x4          -1.12524    0.60239  -1.868   0.0618 .  
---  
Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
    Null deviance: 44.987  on 39  degrees of freedom  
Residual deviance: 29.073  on 37  degrees of freedom  
AIC: 35.073  
  
Number of Fisher Scoring iterations: 6  
回归方程为  
P=exp(-6.13755+0.09759x1-1.12524x4)/(1+exp(-6.13755+0.09759x1-1.12524x4))  
概率估计略。  
  
**Ex6.9**  
我表示不想做了...  
我没弄明白nls()函数里所说的start即初始值怎么设置。好像可以随便设置，只要保证函数收敛即可？？

# 7

Ex7.1

(1)

>lamp<-data.frame(X=c(115,116,98,83,103,107,118,116,73,89,85,97),A=factor(rep(1:3,c(4,4,4))))

> lamp.aov<-aov(X~A,data=lamp);summary(lamp.aov)

Df Sum Sq Mean Sq F value Pr(>F)

A 2 1304 652.0 4.923 0.0359 \*

Residuals 9 1192 132.4

P值小于0.05，有显著差异。

(2)

对甲的区间估计：

> a<-c(115,116,98,83)

> t.test(a)

One Sample t-test

data: a

t = 13.1341, df = 3, p-value = 0.0009534

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

78.04264 127.95736

sample estimates:

mean of x

103

或者用这个命令更简单：

>attach(lamp)

> t.test(X[A==1])

乙的均值估计为111，95%置信区间为99.59932, 122.40068。

丙的均值估计为86，95%置信区间为70.08777, 101.91223。

(3)多重检验：

> attach(lamp)

P值不做调整：

> pairwise.t.test(X,A,p.adjust.method = "none")

Pairwise comparisons using t tests with pooled SD

data: X and A

1 2

2 0.351 -

3 0.066 0.013

P值进行Holm调整：

P value adjustment method: none

> pairwise.t.test(X,A,p.adjust.method = "holm",data)

Pairwise comparisons using t tests with pooled SD

data: X and A

1 2

2 0.35 -

3 0.13 0.04

P value adjustment method: holm

不论采取哪种方法，都可看出乙和丙有显著差异。

Ex7.2

(1)

>lamp<-data.frame(X=c(20,18,18,17,15,16,13,18,22,17,26,19,26,28,23,25,24,25,18,22,27,24,12,14),A=factor(rep(1:4,c(10,6,6,2))))

> lamp.aov<-aov(X ~ A, data=lamp);summary(lamp.aov)

Df Sum Sq Mean Sq F value Pr(>F)

A 3 351.7 117.24 15.11 2.28e-05 \*\*\*

Residuals 20 155.2 7.76

P值小于0.05，可认为四个厂生产的产品的变化率有显著差异。

(2)

> attach(lamp)

P值不做调整：

> pairwise.t.test(X,A,p.adjust.method = "none")

Pairwise comparisons using t tests with pooled SD

data: X and A

1 2 3

2 8.0e-05 - -

3 0.00053 0.47666 -

4 0.05490 6.1e-05 0.00020

P value adjustment method: none

P值进行Holm调整：

> pairwise.t.test(X,A,p.adjust.method = "holm")

Pairwise comparisons using t tests with pooled SD

data: X and A

1 2 3

2 0.00040 - -

3 0.00158 0.47666 -

4 0.10979 0.00036 0.00079

P value adjustment method: holm

由此可得，除了A1和A4，A2和A3这两组的差异不显著外，其他组合的差异都很显著。

Ex7.3

>lamp1<-data.frame(X=c(30,27,35,35,29,33,32,36,26,41,33,31,43,45,53,44,51,53,54,37,47,57,48,42,82,66,66,86,56,52 ,76,83,72,73,59,53),A=factor(rep(1:3,c(12,12,12))))

>attach(lamp1)

正态性检验：

> shapiro.test(X[A==1])

Shapiro-Wilk normality test

data: X[A == 1]

W = 0.9731, p-value = 0.9407

> shapiro.test(X[A==2])

Shapiro-Wilk normality test

data: X[A == 2]

W = 0.9708, p-value = 0.9193

> shapiro.test(X[A==3])

Shapiro-Wilk normality test

data: X[A == 3]

W = 0.9371, p-value = 0.4613

数据在三种水平下均是正态的。

方差齐性检验：

> bartlett.test(X~A,data=lamp1)

Bartlett test of homogeneity of variances

data: X by A

Bartlett's K-squared = 12.139, df = 2, p-value = 0.002312

P值小于0.05，认为各组方差不等。

Ex7.4

>lamp<-data.frame(X=c(2.79,2.69,3.11,3.47,1.77,2.44,2.83,2.52,3.83,3.15,4.70,3.97,2.03,2.87,3.65,5.09,5.41,3.47,4.92,4.07,2.18,3.13,3.77,4.26), g=factor(rep(1:3,c(8,8,8))))

先进行正态性和方差齐性检验以选择使用方差分析aov()还是KW检验kruskal.test()。

正态性检验：

> attach(lamp)

> shapiro.test(X[g==1])

Shapiro-Wilk normality test

data: X[g == 1]

W = 0.9659, p-value = 0.8638

> shapiro.test(X[g==2])

Shapiro-Wilk normality test

data: X[g == 2]

W = 0.983, p-value = 0.9763

> shapiro.test(X[g==3])

Shapiro-Wilk normality test

data: X[g == 3]

W = 0.99, p-value = 0.9951

三组数据都服从正态分布。

方差齐性检验：

> bartlett.test(X~g,data=lamp)

Bartlett test of homogeneity of variances

data: X by g

Bartlett's K-squared = 3.4559, df = 2, p-value = 0.1776

p值大于0.05，可认为三组方差齐。因此选用方差分析aov()或者KW检验kruskal.test()均可。

> kruskal.test(X~g,data=lamp)

Kruskal-Wallis rank sum test

data: X by g

Kruskal-Wallis chi-squared = 7.9322, df = 2, p-value = 0.01895

> lamp.aov<-aov(X~g,data=lamp);summary(lamp.aov)

Df Sum Sq Mean Sq F value Pr(>F)

g 2 6.437 3.218 4.284 0.0275 \*

Residuals 21 15.776 0.751

两种方法得到的p值均小于0.05，可认为3种不同处理的诱导作用不同。

Ex7.5

根据题意，适用配伍组设计的Friedman秩和检验。

>lamp<-data.frame(X=c(23.1,57.6,10.5,23.6,11.9,54.6,21.0,20.3,22.7,53.2,9.7,19.6,13.8,47.1,13.6,23.6,22.5,53.7,10.8,21.1,13.7,39.2,13.7,16.3,22.6,53.1,8.3,21.6,13.3,37.0,14.8,14.8), g=gl(4,8),b=gl(8,1,32)) #其中g表示group，b表示block。

> friedman.test(X~g|b,data=lamp)

Friedman rank sum test

data: X and g and b

Friedman chi-squared = 6.45, df = 3, p-value = 0.09166

P值大于0.05，尚不能拒绝原假设。

Ex7.6

(1)>qua<-data.frame(x=c(4.6,4.3,6.1,6.5,6.8,6.4,6.3,6.7,3.4,3.8,4.0,3.8,4.7,4.3,3.9,3.5,6.5,7.0),a=gl(3,6,18), b=gl(3,2,18))

> qua.aov<-aov(x~a+b+a:b,data=qua);summary(qua.aov)

Df Sum Sq Mean Sq F value Pr(>F)

a 2 3.974 1.987 26.69 0.000164 \*\*\*

b 2 4.441 2.221 29.83 0.000107 \*\*\*

a:b 4 21.159 5.290 71.06 8.34e-07 \*\*\*

Residuals 9 0.670 0.074

两种因素以及其交互作用对产品质量的影响都很显著。

(2)最优条件为A3和B3组合。

> t.test(c(6.5,7.0))

One Sample t-test

data: c(6.5, 7)

t = 27, df = 1, p-value = 0.02357

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

3.573449 9.926551

sample estimates:

mean of x

6.75

点估计为6.75，区间估计为 3.573449，9.926551

(3)双因素方差分析的多重比较？不会...

Ex7.7

正交试验的方差分析。

L9(3^4)正交表。

>pro<-data.frame(Y=c(62.925,57.075,51.6,55.05,58.05,56.55,63.225,50.7,54.45),A=gl(3,3),B=gl(3,1,9),C=factor(c(1,2,3,2,3,1,3,1,2)))

> pro.aov<-aov(Y~A+B+C,data=pro);summary(pro.aov)

Df Sum Sq Mean Sq F value Pr(>F)

A 2 1.76 0.88 0.022 0.978

B 2 65.86 32.93 0.836 0.545

C 2 6.66 3.33 0.085 0.922

Residuals 2 78.78 39.39

结果显示这三个因素对水稻产量的影响均不明显。

K[i,j]<-mean(rate$Y[rate[j]="

> for (j in 1:3)

for (i in 1:3)

K[i,j]<-mean(rate$Y[rate[j]==i])

> K

A B C

1 41 47 45

2 48 55 57

3 61 48 48

B取水平2，A取水平3，C取水平2.

Ex7.8

表示不懂交互作用表。who knows?

Ex7.9

有重复试验的方差分析。

>out<-data.frame(Y=c(1.5,1.7,1.3,1.5,1.0,1.2,1.0,1.0,2.5,2.2,3.2,2.0,2.5,2.5,1.5,2.8,1.5,1.8,1.7,1.5,1.0,2.5,1.3,1.5,1.8,1.5,1.8,2.2,1.9,2.6,2.3,2.0),A=gl(2,16),B=gl(2,8,32),C=gl(2,4,32))

> out.aov<-aov(Y~A+B+C+A:B+A:C+B:C,data=out);summary(out.aov)

Df Sum Sq Mean Sq F value Pr(>F)

A 1 0.008 0.008 0.051 0.8224

B 1 4.728 4.728 31.142 8.37e-06 \*\*\*

C 1 0.038 0.038 0.249 0.6221

A:B 1 1.015 1.015 6.688 0.0159 \*

A:C 1 0.428 0.428 2.818 0.1057

B:C 1 0.263 0.263 1.731 0.2002

Residuals 25 3.795 0.152

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

B的影响最显著，A和B的交互作用影响也很显著。

计算A和B的交互作用：

> ab<-function(x,y){

+ n<-length(x);z<-rep(0,n)

+ for (i in 1:n)

+ if (x[i]==y[i]){z[i]<-1} else{z[i]<-2}

+ factor(z)

+ }

> out$AB<-ab(out$A,out$B)

计算各水平的均值：

> K<-matrix(0,nrow=2,ncol=4,dimnames=list(1:2,c("A","B","C","AB")))

> for (j in 2:5)

+ for (i in 1:2)

+ K[i,j-1]<-mean(out$Y[out[j]==i])

> K

A B C AB

1 1.83750 1.43750 1.85625 1.64375

2 1.80625 2.20625 1.78750 2.00000

由于值越小越好，因此B应用水平1，AB也用水平1，因此A也得用水平1，C用水平2.