Chapter 2 - 均值向量和协方差阵的检验

张笑竹 / 201618070114

2018年10月23日

第二章实验采用课本例2.1中给出的数据，对我国不同行业的上市公司运营情况进行分析。
将本次的实验任务拆分如下：
1）检验3个总体的随机向量是否服从多元正态分布；
2）检验3个总体的随机向量是否具有协方差阵同质性；
3）进行多元方差分析，检验3个总体的均值向量有无显著差异；
4）若3）中所述差异存在，进行一元单因素方差分析，寻找在3个总体间存在差异的具体变量；
5）对4）中找到的存在显著差异的变量，进行事后比较，分析存在差异的具体行业。

1 多元正态分布的检验

1.1 单变量的检验

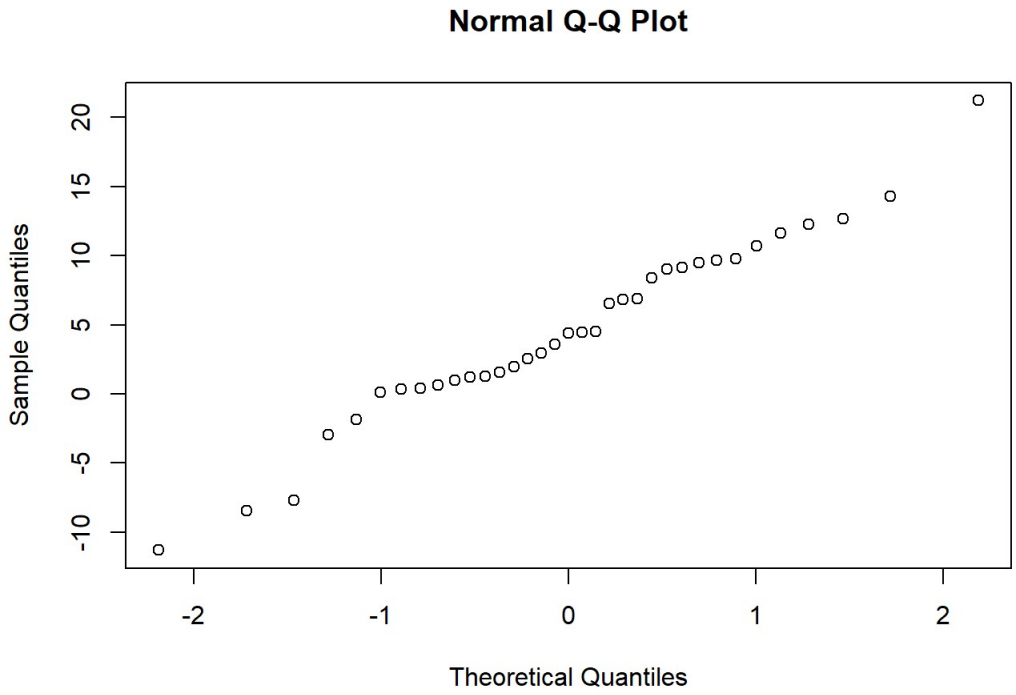
首先，记“净资产收益率”、“总资产报酬率”、“总资产周转率”、“流动资产周转率”、“资产负债率”、“已获利息倍数”、“销售增长率”以及“资本积累率”分别为变量$X_1, X_2, \cdots, X_8$.
将数据复制进入剪贴板，在的R中读入本次实验的数据集。

```
X=read.table("clipboard",header=T)
```

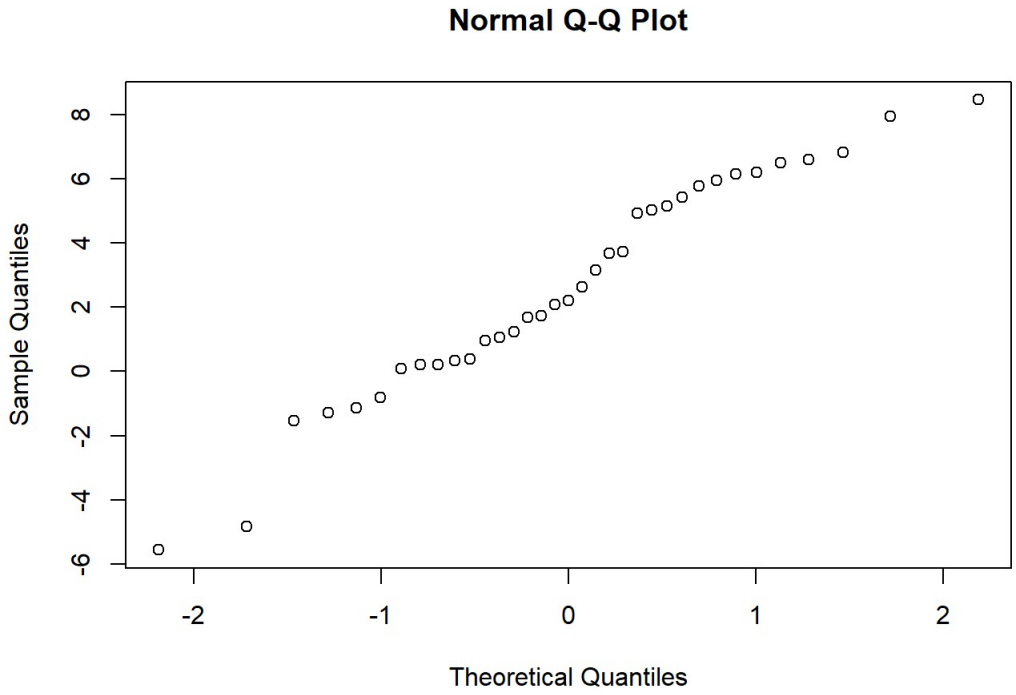
在进行多元方差分析之前，首先要对各个数据是否服从多元正态分布进行检验。
值得注意的是，多元正态分布随机向量的任何一个分量子集的分布仍然服从正态分布；这意味着，如果某一个分量不服从正态分布，那么该多元随机向量必然不服从多元正态分布。因此，为了提高效率，首先，需要对8大指标分别进行正态分布的检验。
由于该例中样本数较少，所以可以利用非参数方法，对数据进行Shapiro检验，并仅针对服从正态分布的变量，输出其检验的p-value和q-q图。

```
library(mvnormtest)
#par(mfrow=c(4,2))
for(i in 4:11){
  k1=mshapiro.test(t(X[,i]))
    if(k1$p.value>0.05){
    print(list(i-3,k1$p.value))
    qqnorm(X[,i])
  }
}
```

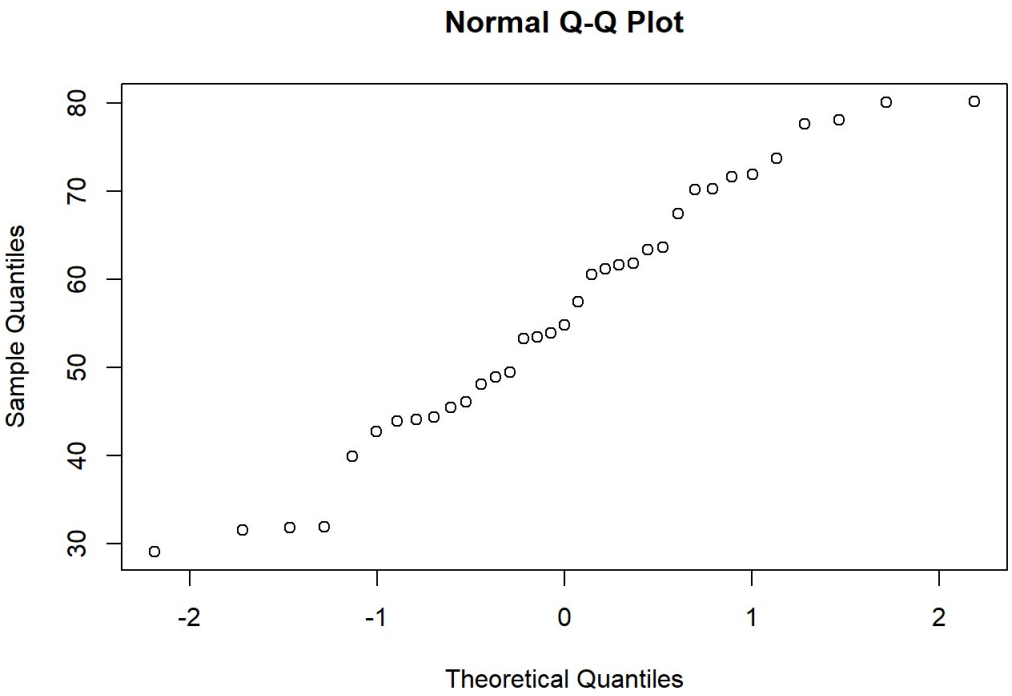
```
## [[1]]
## [1] 1
##
## [[2]]
## [1] 0.6771964
```



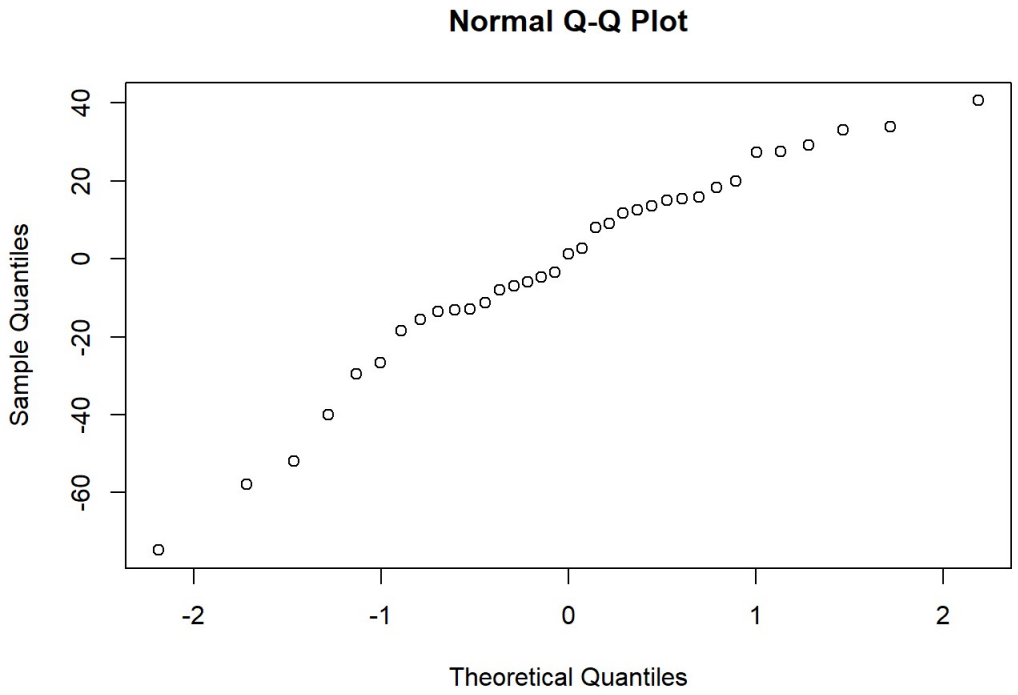

```
## [[1]]
## [1] 2
##
## [[2]]
## [1] 0.2979751
```



```
## [[1]]
## [1] 3
##
## [[2]]
## [1] 0.2653487
```




```
## [[1]]
## [1] 7
##
## [[2]]
## [1] 0.1037941
```



可见，服从一维正态分布的变量仅有上述4个，分别是“净资产收益率”、“总资产报酬率”、“总资产周转率”和“销售增长率”，这与课本中的SPSS输出的结果一致。

1.2 随机向量的检验

此时，已经筛选出4个服从正态分布的一维随机变量。一种自然的想法是，它们组成的随机向量是否服从多元正态分布？利用Shapiro多元检验进行分析。

```
mshapiro.test(t(X[,c(4,5,6,10)]))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.89185, p-value = 0.002405
```

根据检验结果，随机向量(X_1, X_2, X_3, X_7)并不服从正态分布。但是，由于多元正态性检验在SPSS软件中并不容易实现，所以课本案例直接将(X_1, X_2, X_3, X_7)视为多元正态向量；而这显然是不合理的，也可能会影响后续的检验结果。

既然由4个分量组成的随机向量不服从正态分布，就只能退而求其次，寻找由3个分量组成的正态随机向量，一共有$C_4^3 = 4$种选择。

下面筛选出服从正态分布的三维随机向量。

```
#检验3个分量构成的向量是否服从正态分布
for(i in c(4,5,6,10)){
  for(j in c(5,6,10)){
    for(k in c(6,10)){
      if(i<j && j<k){
        te=mshapiro.test(t(X[,c(i,j,k)]))
        if(te$p.value>0.05){
          print(list(c(i-3,j-3,k-3),te$p.value))
        }
      }
    }
  }
}
```

```
## [[1]]
## [1] 1 2 3
##
## [[2]]
## [1] 0.1078618
##
## [[1]]
## [1] 1 3 7
##
## [[2]]
## [1] 0.2530562
```

不难发现，一共有两个三维随机向量服从正态分布。因此，在后续的协方差和均值检验中，应分为两组分别进行检验。

2 协方差阵同质性检验

2.1 函数的封装

首先，按照三大行业，将上述两组正态三维向量进行拆分和重构,共得到6个总体，分别记为(X_1^i, X_2^i, X_3^i)和$(X_1^i, X_3^i, X_7^i), i = 1, 2, 3.$ 然后分别计算出每一个水平的样本数量。

```
#attach(X)
r=3
G1 <- subset(X,行业=="电力、煤气及水的生产和供应业")
G2 <- subset(X,行业=="房地行业")
G3 <- subset(X,行业=="信息技术业")
n1=nrow(G1)
n2=nrow(G2)
n3=nrow(G3)
n=n1+n2+n3
```

然后，利用课本提供的理论，编写函数cov.test()，使得当输入“检验变量个数”和“检验变量编号（向量）”时，可以输出对应的统计量M，以及b和p-value.

```
cov.test <- function(p,nr){ #输入c(检验变量个数,检验变量编号(向量))
  nr=nr+3
  L1=(n1-1)*cov(G1[,nr])
  L2=(n2-1)*cov(G2[,nr])
  L3=(n3-1)*cov(G3[,nr])
  L=L1+L2+L3

  #构造M统计量
  A=(n-r)*log(det(L/(n-r)))
  B1=(n1-1)*log(det(L1/(n1-1)))
  B2=(n2-1)*log(det(L2/(n2-1)))
  B3=(n3-1)*log(det(L3/(n3-1)))
  M=A-B1-B2-B3
  #寻找分布
  if(n1==n2 && n2==n3){
    d1=(2*p^2+3*p-1)*(r-1)/(6*(p+1)*r*(n-1))
    d2=(p-1)*(p+2)*(r^2+r+1)/(6*r^2*(n-1)^2)
  }

  if(n1!=n2 || n1!=n3 || n2!=n3){
    d1=(2*p^2+3*p-1)/(6*(p+1)*(r-1))*(1/(n1-1)+1/(n2-1)+1/(n3-1)-1/(n-r))
    d2=(p-1)*(p+2)/(6*(r-1))*(1/(n1-1)^2+1/(n2-1)^2+1/(n3-1)^2-1/(n-r)^2)
  }

  f1=p*(p+1)*(r-1)/2
  f2=(f1+2)/(d2-d1^2)
  b=f1/(1-d1-f1/f2)
  FS=M/b
  pvalue=pf(FS,f1,f2,ncp=0,lower.tail=FALSE,log.p=FALSE)
  list("M"=M,"b"=b,"p-value"=pvalue)
}
```

2.2 进行检验

调用函数cov.test()，对1.2中筛选出的两个三维正态向量，在不同的行业上，进行协方差阵的齐性检验。首先对(X_1^i, X_2^i, X_3^i)，$i=1,2,3$进行检验：

```
cov.test(3,c(1,2,3))
```

```
## $M
## [1] 15.9962
##
## $b
## [1] 14.07136
##
## $`p-value`
## [1] 0.3246792
```

p-value大于0.05，故保留原假设，认为三个总体(X_1^i, X_2^i, X_3^i)，$i=1,2,3$的协方差阵并无显著差异。
然后，对(X_1^i, X_3^i, X_7^i)，$i=1,2,3$进行检验：

```
cov.test(3,c(1,3,7))
```

```
## $M
## [1] 21.33892
##
## $b
## [1] 14.07136
##
## $`p-value`
## [1] 0.1104601
```

同理，p-value大于0.05，故保留原假设，认为三个总体(X_1^i, X_3^i, X_7^i)，$i=1,2,3$的协方差阵并无显著差异。

3 方差分析

3.1 多元方差分析

经过上述检验，向量(X_1^i, X_2^i, X_3^i)和$(X_1^i, X_3^i, X_7^i), i=1,2,3$均通过了正态性和方差齐性的检验，所以可以进行多元方差分析。

首先，输出各个变量在三大行业间的均值，进行预判断。

```
attach(X)
aggregate(cbind(净资产收益率,总资产报酬率,资产负债率,销售增长率),by=list(行业),FUN=mean)
```

```
##                          Group.1 净资产收益率 总资产报酬率 资产负债率
## 1 电力、煤气及水的生产和供应业     0.1690909    0.5236364   60.31455
## 2                       房地行业     6.8713333    3.5366667   54.84667
## 3                     信息技术业     5.8177778    3.5933333   53.05556
##   销售增长率
## 1  -1.038182
## 2 -10.512000
## 3  12.184444
```

可以大致看出，除了资产负债率X_3，其他三个变量的均值均可能存在行业间的较为明显的差异。为了进行更为准确的判断，进行方差分析。这里采用了Wilks统计量，首先对(X_1^i, X_2^i, X_3^i)进行分析。

```
y.1 <- as.data.frame(cbind(行业,净资产收益率,总资产报酬率,资产负债率))
type1 <- as.factor(y.1$行业)
fit1 <- manova(cbind(净资产收益率,总资产报酬率,资产负债率)~type1,data=y.1)
summary(fit1,test=c("Wilks"))
```

```
##           Df   Wilks approx F num Df den Df  Pr(>F)
## type1      2 0.69829   1.9669      6     60 0.08464 .
## Residuals 32
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

然后对向量(X_1^i, X_3^i, X_7^i)进行分析。

```
y.2 <- as.data.frame(cbind(行业,净资产收益率,资产负债率,销售增长率))
type2 <- as.factor(y.2$行业)
fit2 <- manova(cbind(净资产收益率,资产负债率,销售增长率)~type2,data=y.2)
summary(fit2,test=c("Wilks"))
```

```
##           Df   Wilks approx F num Df den Df  Pr(>F)
## type2      2 0.59849   2.9262      6     60 0.01432 *
## Residuals 32
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

综上，前者的p-value为0.085，而后者的p-value为0.014.
这意味着，当取显著性水平$\alpha = 0.05$时，向量$(X_1^i, X_3^i, X_7^i), i = 1,2,3$的均值存在着显著的差异，但是向量$(X_1^i, X_2^i, X_3^i), i = 1,2,3$的均值差异并不显著。当取显著性水平$\alpha = 0.10$时，向量$(X_1^i, X_2^i, X_3^i), i = 1,2,3$和向量$(X_1^i, X_3^i, X_7^i), i = 1,2,3$的均值均存在着显著的差异。鉴于上述两种情况，无法得到一个确切的结论。因此，在并不严格的情况下，暂且拒绝原假设，认为两个向量的均值在三大行业中不同。

3.2 单因素方差分析

根据上述分析，净资产收益率、总资产报酬率、资产负债率和销售增长率的均值向量在三大行业之间存在显著差异。那么，具体是哪些变量导致了这种差异？这就需要进行单因素方差分析进行检验。

```
y.3 <- as.data.frame(cbind(行业,净资产收益率,总资产报酬率,资产负债率,销售增长率))
type3 <- as.factor(y.3$行业)
fit3 <- manova(cbind(净资产收益率,总资产报酬率,资产负债率,销售增长率)~type3)
summary.aov(fit3)
```

```
##  Response 净资产收益率 :
##             Df Sum Sq Mean Sq F value  Pr(>F)
## type3        2  306.3 153.150  4.0005 0.02814 *
## Residuals   32 1225.0  38.283
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response 总资产报酬率 :
##             Df Sum Sq Mean Sq F value  Pr(>F)
## type3        2  69.46  34.732  3.3201 0.04895 *
## Residuals   32 334.75  10.461
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response 资产负债率 :
##             Df Sum Sq Mean Sq F value Pr(>F)
## type3        2  302.4  151.18  0.6802 0.5137
## Residuals   32 7112.4  222.26
##
##  Response 销售增长率 :
##             Df  Sum Sq Mean Sq F value Pr(>F)
## type3        2  2904.6 1452.29  2.1536 0.1326
## Residuals   32 21579.5  674.36
```

根据和p-value可以判断，均值在三大行业间的差异是由“净资产收益率”X_1和“总资产报酬率”X_2带来的；而“资产负债率”X_3和“销售增长率”X_4的均值在三大行业之间无明显差异。

3.3 事后比较

最后，我们更加关心的是，对于变量X_1和X_2，均值的差异究竟体现在哪些行业之间。一种自然的思路是，进行事后比较。目前常用的方法有SNK检验、Tukey检验、Scheffe检验和Bonferroni检验。此处自定义函数, 利用这4种方法综合分析，以求得到更完善的结论。

```
library(agricolae)
```

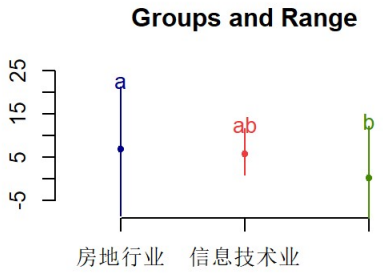
```
## Warning: package 'agricolae' was built under R version 3.4.4
```

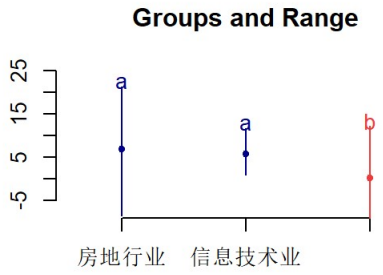
```
back.test <- function(i){
  attach(X)
  fit <- aov(i~行业,data=X)
  #bonferroni检验
  out.LSD <- LSD.test(fit,"行业",p.adj="bonferroni")
  #SNK检验
  out.SNK <- SNK.test(fit,"行业")
  #TukeyHSD检验
  out.TUK=TukeyHSD(fit)
  #Scheffe检验
  out.SHF <- scheffe.test(fit,"行业")

  par(mfrow=c(2,2))
  plot(out.LSD)
  plot(out.SNK)
  plot(out.TUK)
  plot(out.SHF)
  list(LSD=out.LSD$group,SNK=out.SNK$group,TUK=out.TUK$行业,SHF=out.SHF$group)
}
```

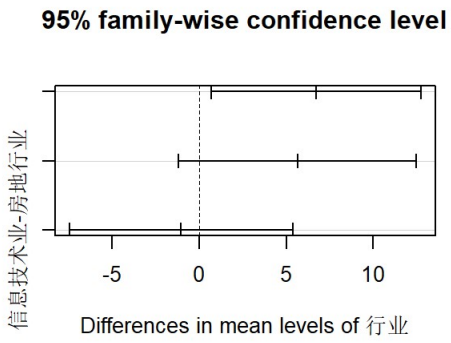
然后，调用该函数back.test(),首先对“净资产收益率”X_1进行检验。

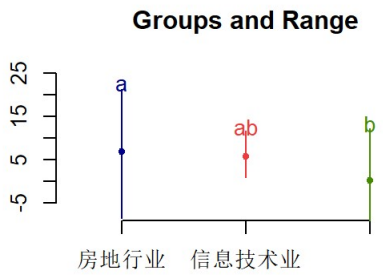
```
back.test(净资产收益率)
```

```
## The following objects are masked from X (pos = 4):
##
##      公司简称, 股票代码, 行业, 净资产收益率, 流动资产周转率,
##      销售增长率, 已获利息倍数, 资本积累率, 资产负债率,
##      总资产报酬率, 总资产周转率
```









```
## $LSD
##                                       i groups
## 房地行业                          6.8713333      a
## 信息技术业                        5.8177778     ab
## 电力、煤气及水的生产和供应业 0.1690909      b
##
## $SNK
##                                       i groups
## 房地行业                          6.8713333      a
## 信息技术业                        5.8177778      a
## 电力、煤气及水的生产和供应业 0.1690909      b
##
## $TUK
##                                                diff        lwr       upr
## 房地行业-电力、煤气及水的生产和供应业     6.702242  0.6666741 12.737811
## 信息技术业-电力、煤气及水的生产和供应业  5.648687 -1.1852485 12.482622
## 信息技术业-房地行业                        -1.053556 -7.4643552  5.357244
##                                                p adj
## 房地行业-电力、煤气及水的生产和供应业    0.02693113
## 信息技术业-电力、煤气及水的生产和供应业 0.12106437
## 信息技术业-房地行业                        0.91426664
##
## $SHF
##                                       i groups
## 房地行业                          6.8713333      a
## 信息技术业                        5.8177778     ab
## 电力、煤气及水的生产和供应业 0.1690909      b
```

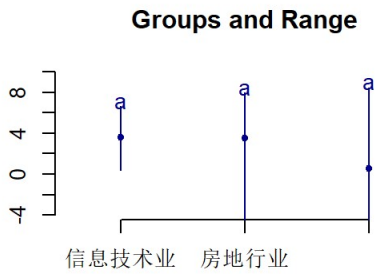
综合考虑上述四种方法给出的结果，可以得到结论，对于变量“净资产收益率”X_1而言，“电力、煤气及水的供应业”分别与“信息技术业”、“房地行业”存在较为显著的差异，而“信息技术业”和“房地行业”之间的差异则并不显著。“房地行业”和“信息技术业”的“净资产收益率”均值比“电力、煤气及水的供应业”显著地高，但是“房地行业”的“净资产收益率”波动更大。

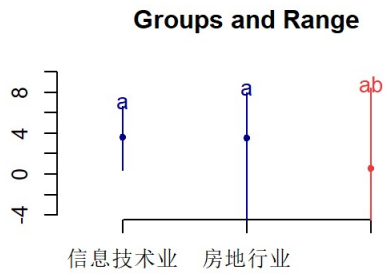
再对“总资产报酬率”X_2进行检验。

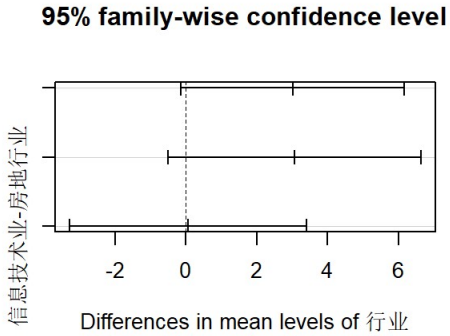
```
back.test(总资产报酬率)
```

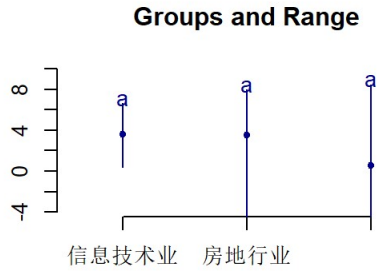
```
## The following objects are masked from X (pos = 3):
##
##     公司简称, 股票代码, 行业, 净资产收益率, 流动资产周转率,
##     销售增长率, 已获利息倍数, 资本积累率, 资产负债率,
##     总资产报酬率, 总资产周转率
```

```
## The following objects are masked from X (pos = 5):
##
##     公司简称, 股票代码, 行业, 净资产收益率, 流动资产周转率,
##     销售增长率, 已获利息倍数, 资本积累率, 资产负债率,
##     总资产报酬率, 总资产周转率
```










```
## $LSD
##                                       i groups
## 信息技术业                     3.5933333      a
## 房地行业                       3.5366667      a
## 电力、煤气及水的生产和供应业 0.5236364      a
##
## $SNK
##                                       i groups
## 信息技术业                     3.5933333      a
## 房地行业                       3.5366667      a
## 电力、煤气及水的生产和供应业 0.5236364     ab
##
## $TUK
##                                              diff        lwr      upr
## 房地行业-电力、煤气及水的生产和供应业    3.01303030 -0.1419955 6.168056
## 信息技术业-电力、煤气及水的生产和供应业 3.06969697 -0.5026663 6.642060
## 信息技术业-房地行业                      0.05666667 -3.2945071 3.407840
##                                             p adj
## 房地行业-电力、煤气及水的生产和供应业    0.06358573
## 信息技术业-电力、煤气及水的生产和供应业 0.10338557
## 信息技术业-房地行业                      0.99904854
##
## $SHF
##                                       i groups
## 信息技术业                     3.5933333      a
## 房地行业                       3.5366667      a
## 电力、煤气及水的生产和供应业 0.5236364      a
```

对于变量“总资产报酬率”X_2而言，“电力、煤气及水的供应业”分别与“信息技术业”、“房地行业”存在一些差异，但是这种差异没有X_1来的明显。而“信息技术业”和“房地行业”之间的差异则并不显著。此外，“房地行业”和“电力、煤气及水的供应业”的“总资产报酬率”比“信息技术业”的波动要大一些。