| 班级 | 信息 1805 | 学号 | 201810000502 | 姓名 | 崔润锋 |
|---|---|---|---|---|---|

结合课程上机 (R 语言), 完成以下上机作业问题:

**注 1** 基本要求:

1) 针对题目要求给出解答, 给出核心关键代码，不必要粘贴所有源代码):

2) 简要概述你通过编程解决此问题所遇到的难点及收获的 R 语言或课程理论等方面的心得。

**注 2** 评价依据:

1) 解答的完整性、正确性: 是否缺少内容、是否计算无误;

2) 难点分析: 是否记录了解决问题过程中的难点和心得;

3) 核心关键代码是否有恰当的注释;

4) 雷同抄袭等: 超期作业评价不超过 60 分, 雷同抄袭一律 0 分。

## 习题 4.5

解: (1) 利用 cor 函数求得相关系数矩阵如下:

| | X1 | X2 | X3 | X4 | X5 |
|---|---|---|---|---|---|
| X1 | 1.00000000 | 0.33364671 | -0.05453868 | -0.06125369 | -0.28936059 |
| X2 | 0.33364671 | 1.00000000 | -0.02290183 | 0.39893102 | -0.15630387 |
| X3 | -0.05453868 | -0.02290183 | 1.00000000 | 0.53332919 | 0.49676279 |
| X4 | -0.06125369 | 0.39893102 | 0.53332919 | 1.00000000 | 0.69842442 |
| X5 | -0.28936059 | -0.15630387 | 0.49676279 | 0.69842442 | 1.00000000 |
| X6 | 0.19879627 | 0.71113407 | 0.03282961 | 0.46791730 | 0.28012915 |
| X7 | 0.34869847 | 0.41359462 | -0.13908580 | -0.17127417 | -0.20827738 |
| X8 | 0.31867736 | 0.83495175 | -0.25835810 | 0.31275728 | -0.08123414 |

| | X6 | X7 | X8 |
|---|---|---|---|
| X1 | 0.19879627 | 0.3486985 | 0.31867736 |
| X2 | 0.71113407 | 0.4135946 | 0.83495175 |
| X3 | 0.03282961 | -0.1390858 | -0.25835810 |
| X4 | 0.46791730 | -0.1712742 | 0.31275728 |
| X5 | 0.28012915 | -0.2082774 | -0.08123414 |
| X6 | 1.00000000 | 0.4168213 | 0.70158588 |
| X7 | 0.41682128 | 1.0000000 | 0.39886792 |
| X8 | 0.70158588 | 0.3988679 | 1.00000000 |

(2) 从样本相关系数矩阵 $R$ 出发做主成分分析，使用 princomp 函数，求得的各主成分的贡献率如图 1所示:

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
Standard deviation     1.759627 1.5385783 0.9591597 0.84019364 0.70600448
Proportion of Variance 0.387036 0.2959029 0.1149984 0.08824067 0.06230529
Cumulative Proportion  0.387036 0.6829389 0.7979373 0.88617801 0.94848330
                            Comp.6     Comp.7      Comp.8
Standard deviation     0.47946669 0.36162932 0.226868982
Proportion of Variance 0.02873604 0.01634697 0.006433692
Cumulative Proportion  0.97721934 0.99356631 1.000000000
```

图 1: 各主成分的贡献率

从中可以看出前两个主成分的累计贡献率为 0.6829389.

(3) 使用 eigen 函数可以求得相关系数矩阵 $R$ 的正交化单位向量，从而前两个主成分分别为:

$$\begin{aligned} Y_1 &= 0.2496x_1 + 0.5192x_2 - 0.0185x_3 + 0.2541x_4 \\ &\quad + 0.02169x_5 + 0.6927x_6 + 0.3171x_7 + 0.5093x_8 \\ Y_2 &= -0.2413x_1 - 0.0376x_2 + 0.4754x_3 + 0.5381x_4 \\ &\quad + 0.5754x_5 + 0.1247x_6 - 0.2607x_7 - 0.0871x_8. \end{aligned}$$

由于 $Y_1$ 和 $Y_2$ 是八个标准化指标的加权值，因此它们反映了平均消费数据的综合指标。对于 $Y_1$, 它反映了各省人均消费水平，除烟茶酒外，其他支出越高，其人均总体消费水平越高，而烟茶酒对其消费水平评价成反方向。在 $Y_2$ 中人均粮食，人均副食品，人均燃料，人均非商品的系数为负；人均烟茶酒、人均其他副食、人均衣着、人均日用品系数为正，说明 $Y_2$ 的绝对值越大，各省人均消费在生活必需品和高档品差异越大。

按照第一主成分将 30 个省、区、市排序，代码如下:

```
# 按第一主成分得分将30个省区市排序
scor <- pca$scores[, 1:2]
print_scor <- scor[(order(scor[,1], decreasing = TRUE)),]
print(print_scor)
```

排序的结果如表 1所示

| 排名 | 省 (区、市) | 第一主成分 | 排名 | 省 (区、市) | 第一主成分 |
|---|---|---|---|---|---|
| 1 | 广东 | 7.01379233 | 16 | 宁夏 | -0.43775323 |
| 2 | 上海 | 3.30394931 | 17 | 湖南 | -0.52687091 |
| 3 | 北京 | 1.82277984 | 18 | 陕西 | -0.62321145 |
| 4 | 浙江 | 1.54097092 | 19 | 云南 | -0.67809647 |
| 5 | 海南 | 1.42511451 | 20 | 新疆 | -0.83249622 |
| 6 | 福建 | 1.17362616 | 21 | 青海 | -1.13238706 |
| 7 | 广西 | 1.07457604 | 22 | 安徽 | -1.13401837 |
| 8 | 天津 | 0.44287310 | 23 | 甘肃 | -1.20243857 |
| 9 | 江苏 | 0.15591226 | 24 | 内蒙古 | -1.27970296 |
| 10 | 辽宁 | 0.04597426 | 25 | 贵州 | -1.28087147 |
| 11 | 西藏 | -0.13551813 | 26 | 吉林 | -1.31581695 |
| 12 | 四川 | -0.13719329 | 27 | 黑龙江 | -1.34833462 |
| 13 | 山东 | -0.14352770 | 28 | 河南 | -1.51135274 |
| 14 | 湖北 | -0.17335839 | 29 | 山西 | -1.71328041 |
| 15 | 河北 | -0.39890334 | 30 | 江西 | -1.99443644 |

表 1: 30 个省、市、区按第一主成分得分排序结果

**习题 4.6**

解: 分别利用 cov 函数和 cor 函数求得样本协方差矩阵为

$$S = \begin{pmatrix} 97.33333 & 17.809524 & 12.02976 & 58.720238 & 22.35119 & 61.529762 \\ 17.80952 & 74.579932 & 14.21854 & 3.326105 & 61.62160 & -3.855867 \\ 12.02976 & 14.218537 & 76.96939 & 41.667517 & 31.21854 & 66.109269 \\ 58.72024 & 3.326105 & 41.66752 & 779.153912 & 310.15944 & 192.423469 \\ 22.35119 & 61.621599 & 31.21854 & 310.159439 & 510.07993 & 156.185799 \\ 61.52976 & -3.855867 & 66.10927 & 192.423469 & 156.18580 & 485.332483 \end{pmatrix}$$

样本相关系数矩阵为

$$R = \begin{pmatrix} 1.0000000 & 0.20903091 & 0.1389848 & 0.21322856 & 0.1003115 & 0.28309667 \\ 0.2090309 & 1.00000000 & 0.1876657 & 0.01379791 & 0.3159387 & -0.02026709 \\ 0.1389848 & 0.18766571 & 1.0000000 & 0.17014805 & 0.1575558 & 0.34204532 \\ 0.2132286 & 0.01379791 & 0.1701481 & 1.00000000 & 0.4919877 & 0.31291525 \\ 0.1003115 & 0.31593872 & 0.1575558 & 0.49198771 & 1.0000000 & 0.31390826 \\ 0.2830967 & -0.02026709 & 0.3420453 & 0.31291525 & 0.3139083 & 1.00000000 \end{pmatrix}$$

基于样本协方差矩阵作主成分分析，结果如图 2所示：

```
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     32.7872275 19.7452189 17.5131159 9.88479688 8.28776305
Proportion of Variance  0.5423404  0.1966919  0.1547353 0.04929446 0.03465271
Cumulative Proportion   0.5423404  0.7390323  0.8937676 0.94306209 0.97771481
                           Comp.6
Standard deviation     6.64625362
Proportion of Variance 0.02228519
Cumulative Proportion  1.00000000
```

图 2: 基于样本协方差矩阵的各主成分贡献率

从图 2可以看出前三个主成分贡献率已占 89.34%, 大于剩下三个主成分的总和，已包含原始数据的大量信息，所以保留前三个主成分即可。

基于相关系数矩阵作主成分分析，结果如图 3所示:

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
Standard deviation     1.4565616 1.0412531 0.9989804 0.9336374 0.76120123
Proportion of Variance 0.3535953 0.1807013 0.1663270 0.1452798 0.09657122
Cumulative Proportion  0.3535953 0.5342966 0.7006236 0.8459034 0.94247458
                           Comp.6
Standard deviation     0.58749684
Proportion of Variance 0.05752542
Cumulative Proportion  1.00000000
```

图 3: 基于相关系数矩阵的各主成分贡献率

从图 3可以看出，前四个主成分已占 84.59% 且第四个主成分的贡献率占到总信息量的 14.53%, 与剩下两个成分的总和差不多，所以保留前四个主成分即可。

我认为基于协方差矩阵 $S$ 的分析结果更合理。因为由协方差矩阵 $S$ 输出结果可以看出前三个主成分的贡献率就可达到 89.38% 大于相关系数矩阵 $R$ 经过主成分分析得到的四个主成分分析贡献率总和 84.59%, 且空腹和摄入食糖的测量数据量纲相等无需进行标准化数据，所以基于协方差矩阵 $S$ 的分析结果更为合理。

## 习题 4.9

解: 首先求得样本协方差矩阵和样本相关系数矩阵分别为

$$\Sigma = \begin{pmatrix} 95.29333 & 52.86833 & 69.66167 & 46.11167 \\ 52.86833 & 54.36000 & 51.31167 & 35.05333 \\ 69.66167 & 51.31167 & 100.80667 & 56.54000 \\ 46.11167 & 35.05333 & 56.54000 & 45.02333 \end{pmatrix}$$

$$R=\begin{pmatrix}1.0000000 & 0.7345555 & 0.7107518 & 0.7039807\\ 0.7345555 & 1.0000000 & 0.6931573 & 0.7085504\\ 0.7107518 & 0.6931573 & 1.0000000 & 0.8392519\\ 0.7039807 & 0.7085504 & 0.8392519 & 1.0000000\end{pmatrix}$$

从相关系数矩阵出发，利用 R 中的 cancor 函数进行典型相关分析，得到两对典型变量分别为

$$\begin{cases}V_1=0.1127152X_1+0.1064583X_2,\\ W_1=0.1029701Y_1+0.1098775Y_2,\end{cases}$$

$$\begin{cases}V_2=-0.2789099X_1+0.2813576X_2,\\ W_2=-0.3610078Y_1+0.3589657Y_2.\end{cases}$$

两对典型变量的典型相关系数分别为

$$\hat{\rho}_1=0.788507916294635,\quad \hat{\rho}_2=0.0537397044242769.$$

从样本协方差矩阵出发，同样可以得到两对典型变量如下:

$$\begin{cases}V_1=0.01154653X_1+0.01443910X_2,\\ W_1=0.01025573Y_1+0.01637533Y_2,\end{cases}$$

$$\begin{cases}V_2=-0.02857148X_1+0.03816093X_2,\\ W_2=-0.03595605Y_1+0.05349758Y_2.\end{cases}$$

其典型相关系数分别为

$$\hat{\rho}_1=0.788507916294635,\quad \hat{\rho}_2=0.0537397044242769.$$

因为样本中测量的数据的量纲都是相同的，所以无论是从协方差矩阵还是相关系数矩阵出发得到的结果都是一样的。

对典型变量进行显著性检验，代码实现如下:

```
corcoef.test <- function(cor, n, p, q) {
    ev <- cor^2
    ev2 <- 1 - ev

    l <- length(ev)
    m <- n - 1 - (p+q+1) / 2
    w <- cbind(NULL)  # 保存中间计算值

```

```
    for (i in 1:l) {
      w <- cbind(w,prod(ev2[i:l]))
    }
Q <- c(NULL)      d <- c(NULL)
    for (i in 1:l){
      Q <- cbind(Q, -(m-(i-1)) * log(w[i]))
      d <- cbind(d, (p-i+1) * (q-i+1))
    }
    # 计算卡方统计量对应的概率
    pvalue <- pchisq(Q, d, lower.tail=FALSE)

    bat <- cbind(t(Q), t(d), t(pvalue))
    colnames(bat) <- c("Chi-Squared", "Num DF", "Pr>F")
    rownames(bat) <- 1:l

    bat     # ret
}
corcoef.test(cca_2 $ cor, n = 25, p = 2, q = 2)
```

运行的结果如表 2所示:

| | Chi-Squared | Num DF | Pr>F |
|---|---|---|---|
| 1 | 20.96417998 | 4 | 0.0003218897 |
| 2 | 0.05928875 | 1 | 0.8076236560 |

表 2: 两个典型变量的显著性检验

取显著水平为 0.05, 其中第一对典型变量的检验 p 值为 0.0003218897, 小于 0.05, 所以第一对典型变量显著相关。而第二对典型变量的检验 p 值为 0.8076236560，大于 0.05, 所以第二对典型变量不是显著相关。

## 习题 5.4

解: 假设两总体服从正态分布且在两总体协方差矩阵相同，即 $\Sigma_1 = \Sigma_2$, 先验概率按比例分配且误判损失相同的条件下进行 Bayes 判别。导入 MASS 包，然后利用其中的 lda 函数进行判别，lda 函数默认就是按比例分配的条件下进行判别的，得到结果如图 4所示:

```
Call:
lda(group ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data = data_frame)

Prior probabilities of groups:
        1         2
0.3428571 0.6571429

Group means:
        x1       x2       x3      x4        x5        x6       x7
1 7.358333 73.66667 1.458333 6.00000 15.250000 0.1716667 49.50000
2 7.686957 67.43478 2.043478 5.23913  6.347826 0.2156522 70.34783

Coefficients of linear discriminants:
             LD1
x1 -1.747293e-01
x2  1.416021e-05
x3  2.072826e-01
x4 -2.052228e-01
x5 -1.935168e-01
x6  8.395266e+00
x7  1.917129e-02
```

图 4: 按比例分配条件下 Bayes 判别结果

从而可以得到线性判别函数为

$$W = -0.17472934x_1 + 0.00001416x_2 + 0.20728259x_3 - 0.20522277x_4 - 0.1935168x_5 + 8.395266x_6 + 0.01917129x_7.$$

下面查看误判情况，使用 table 函数打印出误判情况如表 3所示:

| Group | G1 | G2 |
|---|---|---|
| G1 | 11 | 1 |
| G2 | 1 | 22 |

表 3: 误判情况

另外，还可以打印出误判的详细情况如表 4所示:

| | class | posterior.1 | posterior.2 | LD1 | data.group |
|---|---|---|---|---|---|
| 9 | 2 | 0.2119048 | 0.7880952 | -0.1816424 | 1 |
| 29 | 1 | 0.7578938 | 0.2421062 | -1.0868825 | 2 |

表 4: 误判详细信息

由表 4可以看出第 9 号样本被误判为 G1, 第 29 样本被误判为 G2. 容易求出误判率为 5.71%.

进行交叉确认判别，其结果如表 5所示:

| Group | G1 | G2 |
|---|---|---|
| G1 | 8 | 4 |
| G2 | 2 | 21 |

表 5: 按比例分配条件下交叉确认误判结果

可以计算出误判率为 17.14%.

假设两总体服从正态分布且在两总体协方差矩阵相同，即 $\Sigma_1 = \Sigma_2$, 先验概率按相同的条件下进行 Bayes 判别。同样使用 lda 函数，但是需要指定先验概率相等。代码如下，结果如图 5所示。

```
# 先验概率相等
ld<-lda(group~x1+x2+x3+x4+x5+x6+x7,data=data_frame,prior=c(1,1)/2)
print(ld)
```

```
Call:
lda(group ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data = data_frame,
    prior = c(1, 1)/2)

Prior probabilities of groups:
  1   2
0.5 0.5

Group means:
        x1       x2       x3      x4        x5        x6       x7
1 7.358333 73.66667 1.458333 6.00000 15.250000 0.1716667 49.50000
2 7.686957 67.43478 2.043478 5.23913  6.347826 0.2156522 70.34783

Coefficients of linear discriminants:
             LD1
x1 -1.747293e-01
x2  1.416021e-05
x3  2.072826e-01
x4 -2.052228e-01
x5 -1.935168e-01
x6  8.395266e+00
x7  1.917129e-02
```

图 5: 先验概率相等条件下 Bayes 判别结果

从而可以知道先验概率相等条件下的判别函数和按比例分配条件下的判别函数相同，进行和前面相同的分析可以知道误判情况和表 3相同，同样，此时的误判率为 5.41%. 进行交叉确认判别，其结果如表 6所示。容易计算此时的误判率为 11.43%.

| Group | G1 | G2 |
|---|---|---|
| G1 | 11 | 1 |
| G2 | 3 | 20 |

表 6: 先验概率相等条件下交叉确认误判结果

## 习题 5.5

解: 由于 R 语言内置有鸢尾花数据集，所以直接加载这个数据集，但是注意到数值与教材上的数据相差 10 倍，所以原有数据集处理和教材一致后保存为新的数据集进行分析，相应代码如下:

```
# 导入鸢尾花数据集
data(iris)
# 对鸢尾花数据集进行处理，使其和教材上的数据一致
data_frame <- iris
for (i in 1 : 4) {
    data_frame[i] <- iris[i] * 10
}
names(data_frame) <- c('x1', 'x2', 'x3', 'x4', 'Species')
attach(data_frame)
```

(1) 假定各总体协方差矩阵不全相等，则应该使用二次判别，即使用 qda 函数进行判别，然后使用如下代码进行误判率的回代估计和交叉确认估计。

```
#误判概率
pred2 <- predict(qd1)
table(Species, pred2 $ class)
error3 <- data.frame(pred2, Species)[pred2 $ class != Species,]
length(error3 $ class) / length(pred2 $ class)
# 交叉确认误判率
qd2 <- qda(Species ~ x2 + x4, data = data_frame, CV = TRUE)
table(Species, qd2 $ class)
error4 <- data.frame(pred2, Species)[qd2 $ class != Species,]
length(error4 $ class) / length(pred2 $ class)
```

运行的结果如表 7和表 8所示。依次求得误判率为 4.67% 和 5.33%.

| Species | setosa | versicolor | virginica |
|---|---|---|---|
| setosa | 50 | 0 | 0 |
| versicolor | 0 | 46 | 4 |
| virginica | 0 | 3 | 47 |

表 7: 只考虑指标 X2 和 X4 且协方差矩阵不全相等的误判情况

| Species | setosa | versicolor | virginica |
|---|---|---|---|
| setosa | 50 | 0 | 0 |
| versicolor | 0 | 45 | 4 |
| virginica | 0 | 3 | 47 |

表 8: 只考虑指标 X2 和 X4 且协方差矩阵不全相等的交叉确认估计

(2) 假定各总体协方差矩阵相同，这时使用线性判别，即使用 lda 函数进行判别。判别的结果如图 6所示。

```
Call:
lda(Species ~ x2 + x4, data = data_frame, var.equal = TRUE)

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333

Group means:
               x2    x4
setosa     34.28  2.46
versicolor 27.70 13.26
virginica  29.74 20.26

Coefficients of linear discriminants:
          LD1        LD2
x2 -0.1986964 0.26800746
x4  0.5477136 0.08169648

Proportion of trace:
   LD1    LD2
0.9884 0.0116
```

图 6: 只考虑指标 X2 和 X4 且协方差矩阵相等的判别结果

从而可以得到线性判别函数为

$$W = -0.1986964X_2 + 0.5477136X_4.$$

回代估计和交叉确认估计的结果分别见表 9和表 10所示。误判率分别为 3.33% 和 4%.

| Species | setosa | versicolor | virginica |
|---|---|---|---|
| setosa | 50 | 0 | 0 |
| versicolor | 0 | 49 | 1 |
| virginica | 0 | 4 | 46 |

表 9: 只考虑指标 X2 和 X4 且协方差矩阵相等的误判情况

| Species | setosa | versicolor | virginica |
|---|---|---|---|
| setosa | 50 | 0 | 0 |
| versicolor | 0 | 48 | 2 |
| virginica | 0 | 4 | 46 |

表 10: 只考虑指标 X2 和 X4 且协方差矩阵相等的交叉确认估计

(3) 利用前两问的分析结果，使用 predict 函数进行预测新数据应该判为哪一类结果如图 7所示。

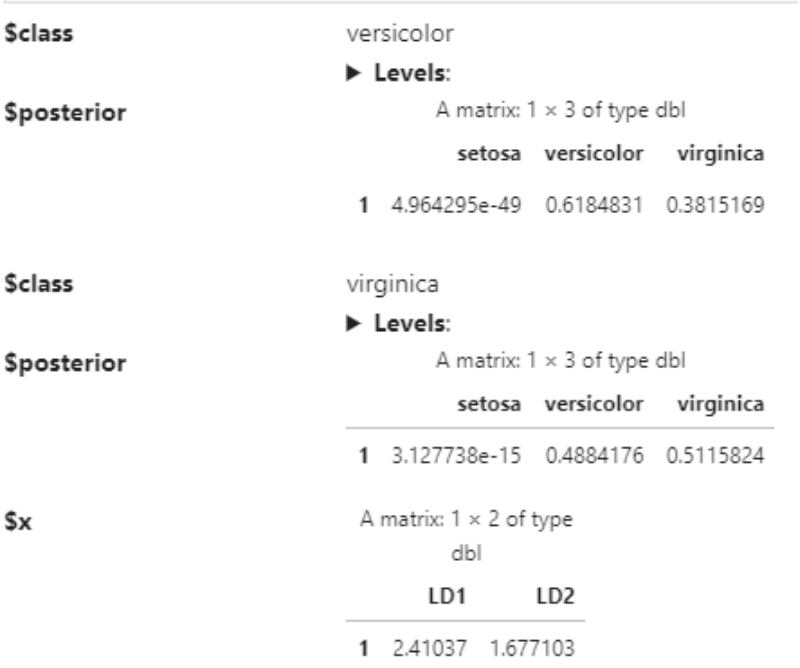


图 7: 新样本的判别归属

由图 7可以知道，在各总体协方差矩阵不全相同的条件下新样品被判别归 versicolor 类；在各总体协方差矩阵相同的条件下，新样品被叛别归 virginica 类。

(4) 在使用全部指标且各总体协方差矩阵不相等的条件下，进行类似第 (1) 问的分析得到的回代估计结果如表 11所示，交叉确认估计的结果如表 12所示。他们的误判率分别为 2% 和 2.67%.

| Species | setosa | versicolor | virginica |
|---|---|---|---|
| setosa | 50 | 0 | 0 |
| versicolor | 0 | 48 | 2 |
| virginica | 0 | 1 | 49 |

表 11: 考虑全部指标且协方差矩阵不相等的误判情况

| Species | setosa | versicolor | virginica |
|---|---|---|---|
| setosa | 50 | 0 | 0 |
| versicolor | 0 | 47 | 3 |
| virginica | 0 | 1 | 49 |

表 12: 考虑全部指标且协方差矩阵不相等的交叉确认估计

在使用全部指标且各总体协方差矩阵相等的条件下，进行类似第 (2) 问的分析得到的回代估计结果和交叉确认估计的结果均如表 13所示，他们的误判率均为 2%.

| Species | setosa | versicolor | virginica |
|---|---|---|---|
| setosa | 50 | 0 | 0 |
| versicolor | 0 | 48 | 2 |
| virginica | 0 | 1 | 49 |

表 13: 考虑全部指标且协方差矩阵相等的误判情况和交叉确认估计

和前两问比较，使用全部指标的误判率要低于仅使用指标 X2 和 X4 的误判率。再分析仅使用指标 X1 和 X3 在各总体协方差矩阵不相等和相等两种情况下的回代估计和交叉确认估计，具体结果这里不再给出，仅写出各自的误判率如表 14所示。

## 习题 6.3

解

| | 回代估计 | 交叉确认估计 |
|---|---|---|
| 协方差矩阵相等 | 3.33% | 4% |
| 协方差矩阵不等 | 4% | 4% |

表 14: 仅考虑指标 X1 和 X3 在不同条件下的回代估计和交叉确认估计的误判率

**习题 6.6**

**习题 6.7**