Homework 4

张子乐,3230104237

处理学习样本

```
df <- read_excel(" 肝胆病患者检查数据.xls",skip = 1)
df <- df[,-1] # 去除第一列
cols <- names(df)
df <- as.data.frame(lapply(df[cols], as.numeric))
df$sick <- ifelse(df$group == 5,0,1) # 只有第五组是正常的
df$sick <- factor(df$sick,levels = c(0, 1),labels = c("Healthy", "Sick")) # 0 为正常, 1 为患病
df <- df[,-5] # 删除 group 列
df1 <- df[df$sick == "Healthy",1:4]
df2 <- df[df$sick == "Sick",1:4]

# 处理测试样本
data <- read_excel(" 体检资料.xls")
data <- data[,c(" 体检编号"," 总胆红素（T-BIL）"," 白蛋白（Alb）"," 碱性磷酸酶（ALP）"," 谷丙转氨酶（A
names(data) <- c("ID", "BIL", "Alb", "ALP", "ALT")
cols <- names(data[,-1])
data[-1] <- as.data.frame(lapply(data[cols], as.numeric))
```

检验均值向量是否有显著差异

H0：健康个体和患病个体四个指标的均值向量没有显著差异

H1：健康个体和患病个体四个指标的均值向量有显著差异

结果显示，霍特林统计量 $T^2 = 78.743$，$p < 0.05$，在 95% 的置信水平下拒绝原假设，选择备择假设，即健康个体和患病个体四个指标的均值向量有显著差异，可以进行判别分析。

```
test1 <- function(df1,df2)
{
  res1 <- HotellingsT2(df1,df2)
```

```
  print(res1)
  return (res1)
}

res1 <- test1(df1,df2)
```

```
##
##  Hotelling's two sample T2-test
##
## data:  df1 and df2
## T.2 = 78.743, df1 = 4, df2 = 339, p-value < 2.2e-16
## alternative hypothesis: true location difference is not equal to c(0,0,0,0)
```

检验组间方差是否有显著差异

H0：两总体的协方差阵相同

H1：两总体的协方差阵不相同

假设检验结果显示，$p < 0.05$，在 95% 的置信水平下拒绝原假设，选择备择假设，因此认为健康总体和患病总体的协方差阵不同。

```
test2 <- function(df)
{
  # s1 <- cov(df1)
  # s2 <- cov(df2)
  # covmat <- list(s1,s2)
  # res2 <- varcomp(covmat,n=c(nrow(df1),nrow(df2)))
  res2 <- boxM(df[, c("BIL", "Alb", "ALP", "ALT")],df$sick) # 似然比检验
  print(res2)
  return(res2)
}
res2 <- test2(df)
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  df[, c("BIL", "Alb", "ALP", "ALT")]
## Chi-Sq (approx.) = 1532.1, df = 10, p-value < 2.2e-16
```

建立判别函数

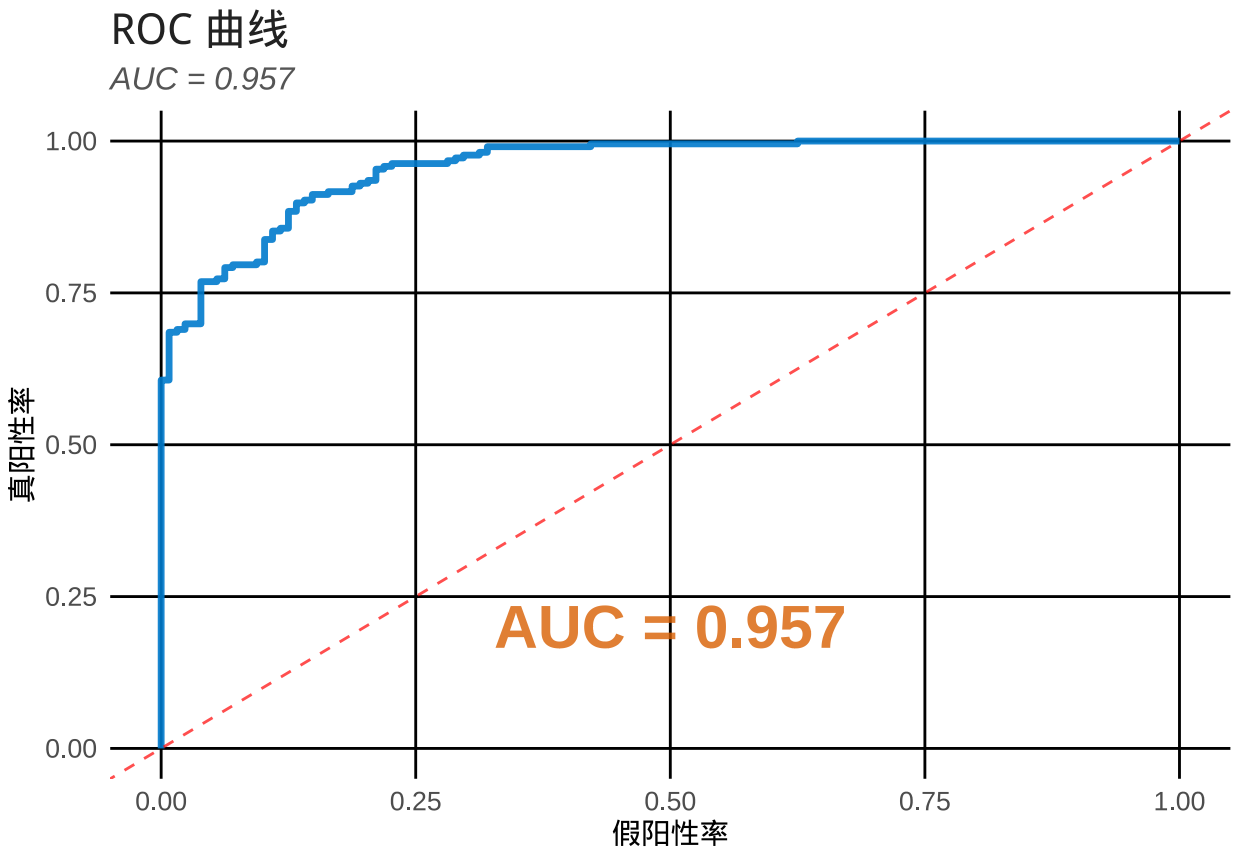
```
test3 <- function(df)
{
  res3 <- qda(sick~BIL+Alb+ALP+ALT,df)
  print(res3)
  return(res3)
}
res3 <- test3(df)
```

```
## Call:
## qda(sick ~ BIL + Alb + ALP + ALT, data = df)
##
## Prior probabilities of groups:
##  Healthy     Sick
## 0.372093 0.627907
##
## Group means:
##               BIL      Alb       ALP       ALT
## Healthy 14.05234 47.69375  60.10938  23.28906
## Sick    64.24167 36.75741 121.27778 262.08333
```

画出 ROC-AUC 曲线图：

```
train_pred <- predict(res3, df[c("BIL", "Alb", "ALP", "ALT")])
pred_roc <- prediction(train_pred$posterior[, "Sick"], df$sick)

perf <- performance(pred_roc, "tpr", "fpr")
auc <- performance(pred_roc, "auc")@y.values[[1]]

roc_data <- data.frame(
  FPR = perf@x.values[[1]],   # 假阳性率
  TPR = perf@y.values[[1]]    # 真阳性率
)

p <- ggplot(roc_data, aes(x = FPR, y = TPR)) +
  geom_line(color = "#007acc", linewidth = 1.2, alpha = 0.9) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "red", alpha = 0.7) +
  labs(
```

```
    title = "ROC 曲线",
    subtitle = paste("AUC =", round(auc, 3)),
    x = " 假阳性率",
    y = " 真阳性率"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 16, face = "bold", color = "#222222"),
    plot.subtitle = element_text(size = 12, face = "italic", color = "#555555"),
    axis.title = element_text(size = 11),
    axis.text = element_text(size = 10),
    panel.grid.major = element_line(color = "black", linewidth = 0.5),
    panel.grid.minor = element_blank(),
    legend.position = "none"
  ) +
  annotate("text", x = 0.5, y = 0.2,
           label = paste("AUC =", round(auc, 3)),
           size = 8, color = "#d95f02", fontface = "bold", alpha = 0.8)

print(p)
```



ROC 曲线靠近左上角，且 $AUC = 0.957 > 0.9$，说明判别效果较好。

根据判别函数分析体检数据

其中 Alb 列有一个缺失值，将包含它的行删除。

```
test_data <- na.omit(data[,c("BIL", "Alb", "ALP", "ALT")])
res4 <- predict(res3,test_data)
print(res4$class)
```

```
##   [1] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy
##  [10] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy
##  [19] Healthy Healthy Sick    Healthy Healthy Healthy Healthy Healthy Healthy
##  [28] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy Sick
##  [37] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy
##  [46] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy
##  [55] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy
##  [64] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy
##  [73] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy
```

```
##  [82] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy
##  [91] Healthy Healthy Healthy Healthy Sick    Healthy Healthy Healthy Healthy
## [100] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy
## [109] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Sick    Healthy
## [118] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy
## [127] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy
## [136] Healthy Healthy Healthy Sick    Healthy Sick    Healthy Healthy Healthy
## [145] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy
## [154] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy
## [163] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy
## [172] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy
## [181] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy
## [190] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy
## [199] Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy Healthy
## [208] Healthy Healthy Healthy Healthy Sick    Healthy Healthy Healthy Healthy
## [217] Healthy Sick    Healthy Healthy Healthy Healthy
## Levels: Healthy Sick
```

```r
prop.table(table(res4$class)) # 计算百分比
```

```
##
##    Healthy       Sick
## 0.96396396 0.03603604
```

222 份体检数据中，96.4% 的个体被判断为健康人，仅 3.6% 的个体被判断为患有肝胆疾病。

```r
means <- t(res3$means) %>%
  as.data.frame() %>%
  tibble::rownames_to_column(var = "Variable") %>%
  rename(Healthy = Healthy, Sick = Sick)

res <- means |>
  mutate(
    `Healthy` = round(Healthy, 2),
    `Sick` = round(Sick, 2)
  ) %>%
  dplyr::select(Variable, Healthy, Sick)

kable(res,
      caption = " 两组各指标的均值",
```

```
    col.names = c(" 变量", " 健康组均值", " 患者组均值"),
    align = "c",
    digits = 2)
```

表 1: 两组各指标的均值

变量	健康组均值	患者组均值
BIL	14.05	64.24
Alb	47.69	36.76
ALP	60.11	121.28
ALT	23.29	262.08

报告

本次实验利用已有医院病人的资料，根据正常人和患者的四项指标（总胆红素 (umol/L)，白蛋白（g/L），碱性磷酸酶，谷丙转氨酶），学习构建了一个判别分析模型，以用于分析未标注是否患病的体检数据。

在构建模型前，对两总体四项指标的均值向量是否有显著差异做假设检验。霍特林 T^2 检验结果显示，健康与患病群体在这些指标的均值上存在显著差异（$p < 0.05$），因此做判别分析是有意义的。同时，对两总体的协方差矩阵作似然比检验，$p < 0.05$，可以认为两总体的协方差矩阵不相等，因此采用广义平方距离 QDA。

对其的效果进行评估：该判别模型具有出色的区分能力，而 ROC 曲线分析得出的 AUC 值高达 0.957，这表明模型在诊断性能上达到了优秀水平，能够高效且准确地识别两类人群。

最后，将此判别模型应用于一份包含 222 个体的体检数据。分析结果显示，其中 96.4% 的个体健康，而约 3.6% 的个体被识别为具有肝胆疾病风险。