Exploratory Multivariate Data Analysis – Hw 3

Cheng-Ying Chou

10/13/2020

Table of Contents

library(tidyverse)

## ─ Attaching packages ────────────────── tidyverse 1.3.0 ─

## ✓ ggplot2 3.2.1 ✓ purrr 0.3.3  
## ✓ tibble 3.0.3 ✓ dplyr 1.0.2  
## ✓ tidyr 1.0.0 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.4.0

## Warning: package 'tibble' was built under R version 3.6.2

## Warning: package 'dplyr' was built under R version 3.6.2

## ─ Conflicts ─────────────────── tidyverse\_conflicts() ─  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(glmnet)

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

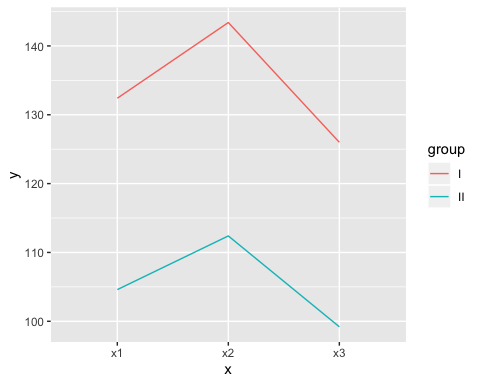
## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

## Loaded glmnet 3.0-1

## Problem 1

In Olkin and Veath (1980), the evolution of citrate concentrations in plasma is observed at three different times of day, (8 am), (11 am) and (3 pm), for two groups of patients who follow different diets. (The patients were randomly attributed to each group under a balanced design .  
Test if the profiles of the groups are parallel, if they are at the same level and if they are horizontal. (plasma.dat)

p1 <- read.table("/Users/linyuxiang/R/MSA/Data/plasma.dat", header = T)  
p1 %>% group\_by(group) %>%   
 summarise\_at(c("x1", "x2", "x3"),mean) %>%   
 gather("x", "y", -"group") %>%   
 ggplot(aes(x = x, y = y, group = group, color = group)) +  
 geom\_line()



g1 <- subset(p1, group=="I")[,-1]; g2 <- subset(p1, group =="II")[,-1]  
S\_g1 <- cov(g1); S\_g2 <- cov(g2); pooled <- (S\_g1+S\_g2)/2  
n1 <- nrow(g1); n2 <- nrow(g2);n <- n1+n2   
x1\_bar <- apply(g2, 2, sum); x2\_bar <- apply(g2, 2, mean)  
x\_bar <- (x1\_bar+x2\_bar)/2  
p <- ncol(p1)-1  
A <- matrix(c(1,-1,0, 0,1,-1),nrow = 2, byrow= T)  
  
# parallelism  
para\_T <- ((n1\*n2) / (n1+n2)) \* t(x1\_bar - x2\_bar) %\*% t(A) %\*%   
 solve(A%\*%pooled%\*%t(A)) %\*% A %\*%(x1\_bar - x2\_bar)  
df1 <- p-1; df2 <- n1+n2-p  
(n1+n2-p) / ((n1+n2-2)\*(p-1)) \* para\_T > qf(0.95, df1,df2)

## [,1]  
## [1,] TRUE

# Equal mean response  
equal\_resp\_T2 <- (n1+n2)\* (t(x\_bar)%\*%t(A))%\*%solve(A%\*%pooled%\*%t(A))%\*%(A%\*%x\_bar)  
df1 <- p-1; df2 <- n1+n2-p  
(n1+n2-p)/((n1+n2-2)\*(p-1)) \* equal\_resp\_T2 > qf(0.95, df1, df2)

## [,1]  
## [1,] TRUE

# same level  
vec\_1 <- matrix(rep(1,length(x\_bar)),nrow = 1)  
abs(vec\_1%\*%(x1\_bar-x2\_bar) /   
 sqrt((vec\_1%\*%pooled%\*%t(vec\_1)) \* (1/n1 + 1/n2))) > qt(0.95,n1+n2-2 )

## [,1]  
## [1,] TRUE

## Problem 2

Baten, Tack, and Baeder (1958) compared judges’ scores on fish prepared by three methods. Twelve fish were cooked by each method, and several judges tasted fish samples and rated each on four variables: y1 = aroma, y2 = flavor, y3 = texture, and y4 =moisture. The data are in the file “fish.csv”. Each entry is an average score for the judges on that fish. Compare the three methods using all four MANOVA tests step by step in addition to using R packages.

p2 <- read.csv("/Users/linyuxiang/R/MSA/Data/fish.csv")  
colnames(p2) <- c("method", "aroma", "flavor", "texture", "moisture")  
  
p <- ncol(p2)-1  
m <- p2$method %>% as.factor() %>% unique %>% length  
n <- nrow(p2)  
## T  
T\_matrix <- matrix(rep(0,p^2), nrow = p)  
for(i in 1:ncol(p2[,-1])){  
 for (j in 1:ncol(p2[,-1])){  
 T\_matrix[i,j] <- (p2[,i+1] %\*% p2[,j+1]) -   
 (sum(p2[,i+1]) \* sum(p2[,j+1]))/nrow(p2)  
   
 }  
}  
## B  
B\_matrix <- matrix(rep(0,p^2), nrow = p)  
temp\_B <- p2 %>%   
 group\_by(method) %>%   
 summarise\_at(c("aroma", "flavor", "texture", "moisture") , sum) %>%   
 as.matrix()  
for(i in 1:ncol(temp\_B[,-1])){  
 for (j in 1:ncol(temp\_B[,-1])){  
 B\_matrix[i,j] <- (temp\_B[,i+1] %\*% temp\_B[,j+1])/(nrow(p2)/m) -   
 (sum(temp\_B[,i+1]) \* sum(temp\_B[,j+1]))/nrow(p2)  
   
 }  
}  
W\_matrix <- T\_matrix - B\_matrix  
lambda <- det(W\_matrix) / det(T\_matrix)  
# Wilks' F  
wilks\_F <- ((1-sqrt(lambda)) / sqrt(lambda)) \* ((n - m - p + 1) / (p))  
df1 <- 2\*p  
df2 <- 2\*(n-m-p+1)  
wilks\_F > qf(0.95,df1,df2)

## [1] TRUE

# Run by package  
y <- cbind(p2$aroma, p2$flavor, p2$texture, p2$moisture)  
mod <- manova(y~as.factor(method), data = p2)  
summary(mod, test = "Wilks")

## Df Wilks approx F num Df den Df Pr(>F)   
## as.factor(method) 2 0.22005 8.4883 8 60 1.223e-07 \*\*\*  
## Residuals 33   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Roy  
lambda <- eigen(solve(W\_matrix) %\*% B\_matrix)$values  
Roy\_lambda <- max(lambda)  
d <- max(c(p, m-1))  
Roy\_F <- ((n - d - 1) \* Roy\_lambda / d ); Roy\_F

## [1] 23.52562

# F test  
df1 <- d; df2 <- n-d-1  
##### 簡報df2寫成 n-m-d-1  
Roy\_F > qf(0.95, df1,df2)

## [1] TRUE

summary(mod, test = "Roy")

## Df Roy approx F num Df den Df Pr(>F)   
## as.factor(method) 2 3.0356 23.526 4 31 5.139e-09 \*\*\*  
## Residuals 33   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Pillai  
Pillai\_V <- sum(lambda/(1+lambda))  
s <- min(c(p, m-1))  
Pillai\_F <- (n-m-p+s) \* Pillai\_V / (d\*(s-Pillai\_V)); Pillai\_F

## [1] 5.896599

# F test  
df1 <- s\*d; df2 <- s\*(n-m-p+s)  
##### 簡報df2 寫成 s\*(n-m-p+s)  
Pillai\_F > qf(0.95, df1, df2)

## [1] TRUE

summary(mod, test = "Pillai")

## Df Pillai approx F num Df den Df Pr(>F)   
## as.factor(method) 2 0.86419 5.8966 8 62 1.32e-05 \*\*\*  
## Residuals 33   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Hotelling-Lawley  
Hotel\_U <- sum(lambda)  
Hotel\_A <- (abs(m-p-1)-1)/2  
Hotel\_B <- (n-m-p-1)/2  
df1 <- s\*(2\*Hotel\_A+s+1); df2 <- 2\*(s\*Hotel\_B+1)  
Hotel\_F <- df2\*Hotel\_U / (s\*df1); Hotel\_F

## [1] 11.46105

Hotel\_F> qf(0.95, df1, df2)

## [1] TRUE

summary(mod, test = "Hotelling-Lawley")

## Df Hotelling-Lawley approx F num Df den Df Pr(>F)   
## as.factor(method) 2 3.1617 11.461 8 58 1.392e-09 \*\*\*  
## Residuals 33   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Problem 3

The file turtle.dat contains the carapace measurements for 24 female and 24 male turtles. Use Box’s M-test to test , where is the population covariance matrix for carapace measurements for female turtles, and is the population covariance matrix for carapace measurements for male turtles. Set . Implement the test step-by-step and compare the result obtained by R package.

library(covTestR)  
p3 <- read.table("/Users/linyuxiang/R/MSA/Data/turtle.dat", header = T)  
f <- subset(p3, female =="female")[,1:3]   
m <- subset(p3, female == "male")[,1:3]  
s\_m <- cov(m); f\_m <- cov(f)  
  
j <- 2; p <- 3; n\_m <- nrow(m); n\_f <- nrow(f);n <- nrow(p3)  
pooled <- ((n\_m/n) \* s\_m ) + ((n\_f/n) \* f\_m )  
Boxes\_M <- ((n\_m-1) \* log(det(pooled)) + (n\_f-1) \* log(det(pooled))) - ((n\_m-1) \* log(det(s\_m)) + (n\_f-1) \* log(det(f\_m)))  
  
Boxes\_C <- 1-( ((2\*(p^2)+3\*p-1)/ (6\*(p+1)\*(j-1))) \* (((1/(n\_m-1))+(1/(n\_f-1))) - ((1/(n\_m+n\_f))\*j)) )  
# chi-test  
  
Boxes\_M\*Boxes\_C > qchisq(0.95, 3)

## [1] TRUE

Boxes\_M\*Boxes\_C

## [1] 22.60577

homogeneityCovariances(p3,group = female)

##   
## Boxes' M Homogeneity of Covariance Matrices Test  
##   
## data: female and male  
## Chi-Squared = 23.754, df = 276, p-value = 1  
## alternative hypothesis: true difference in covariance matrices is not equal to 0

# ###test  
# p3 <- bind\_rows(p3, data\_frame(X98 = 1, X81 = 2, X38 = 3, female = "female"))  
# f <- subset(p3, female =="female")[,1:3]   
# m <- subset(p3, female == "male")[,1:3]  
# s\_m <- cov(m); f\_m <- cov(f)  
#   
# j <- 2; p <- 3; n\_m <- nrow(m); n\_f <- nrow(f);n <- nrow(p3)  
# pooled <- ((n\_m/n) \* s\_m ) + ((n\_f/n) \* f\_m )  
#   
# Boxes\_M <- ((n\_m-1) \* log(det(pooled)) + (n\_f-1) \* log(det(pooled))) - ((n\_m-1) \* log(det(s\_m)) + (n\_f-1) \* log(det(f\_m)))  
# Boxes\_C <- 1- ((2\*(p^2)+3\*p-1)/ (6\*(p+1)\*(j-1))) \* ((2/23) - (1/23))  
# # chi-test  
# Boxes\_M\*Boxes\_C  
#   
# homogeneityCovariances(p3,group = female)