# 安裝R套件

install.packages("formatR")

install.packages("highlight")

install.packages("xtable")

install.packages("ggplot2")

install.packages("dplyr")

install.packages("MASS")

install.packages("ElemStatLearn")

install.packages("HSAUR3")

require(formatR)

require(highlight)

require(xtable)

require(ggplot2)

require(dplyr)

require(MASS)

require(ElemStatLearn)

require(HSAUR3)

# 設定工作目錄

setwd("C:/RData")

# Loads specified data sets, or list the available data sets 由套件HSAUR3裝資料集 heptathlon

data("heptathlon", package = "HSAUR3")

# 印出七項全能heptathlon這個data frame 的第一部分(前6筆)

head(heptathlon)





# transform values consistent 將數值轉換(一般跳高跳遠數字愈大表現愈好, 但跑步類, 是數字愈小表現愈好)

heptathlon$hurdles <- max(heptathlon$hurdles) - heptathlon$hurdles

heptathlon$run200m <- max(heptathlon$run200m) - heptathlon$run200m

heptathlon$run800m <- max(heptathlon$run800m) - heptathlon$run800m

# 找出行變數名稱是score的是哪一行

score <- which(colnames(heptathlon) == "score")



# 去掉第8行(score=8), heptathlon[ , -score]

將heptathlon剩下的變數兩兩相關係數矩陣算出, cor(heptathlon[ , -score])

並只取到小數點第2位, round(cor(heptathlon[ , -score]), 2)

round(cor(heptathlon[ , -score]), 2)



# pair scatter plot # heptathlonPair 畫兩兩散佈圖矩陣

plot(heptathlon[,-score])



# 由圖中發現有一個離群值

# remove outlier 移除離群值PNG(Papua New Guinea巴布亞新幾內亞)

## # grep 是Pattern Matching and Replacement

# 找出heptathlon的列名稱為PNG的有哪些列



#將heptathlon的第25列移除

heptathlon <- heptathlon[-grep("PNG", rownames(heptathlon)),]



# 再重新計算相關係數矩陣

round(cor(heptathlon[ , -score]), 2)



# 下面這指令可有可無

score <- which(colnames(heptathlon) == "score")

#再畫移除離群值的散佈圖矩陣

# pair scatter plot (remove outlier) # heptathlonPair2

plot(heptathlon[,-score])



# 離群值已不在

# PCA 主成分分析做完存到heptathlon\_pca裡

# a logical value indicating whether the variables should be scaled to have unit variance before the analysis takes place

# scale=TRUE代表需要在分析前將變數尺度轉換, 使得變異數為1

heptathlon\_pca <- prcomp(heptathlon[,-score], scale=TRUE)

# 將主成分分析結果(含每個主成分的標準差與負荷量)印出

print(heptathlon\_pca)



#第一個主成分

C1=-0.4503876X1-0.3145115X2-0.4024884X3-0.4270860X4-0.4509639X5-0.2423079X6-0.3029068X7

# 摘要主成分分析結果(含每個主成分的標準差, 可解釋變異比例與累積可解釋變異比例)

summary(heptathlon\_pca)



# 兩個主成分累積可解釋74%的樣本變異

# loading coeff將主成分分析結果的負荷量(係數)印出



# 將第一個主成分分析結果的負荷量(係數)印出

a1 <- heptathlon\_pca$rotation[,1]

a1



# 由係數可知每一項運動的權重都差不多

 所以可將第一個主成分命名為一般運動技能

# 計算每個主成分的估計值

predict(heptathlon\_pca)



# 計算第一個主成分的估計值

predict(heptathlon\_pca)[,1]



# heptathlon[1,] 第一位選手7項運動分數

 hurdles highjump shot run200m longjump javelin run800m score

Joyner-Kersee (USA) 3.73 1.86 15.8 4.05 7.27 45.66 34.92 7291

帶入主成分公式中

-35.27347=-0.4503876\*(3.73)-0.3145115\*(1.86)-0.4024884\*(15.8)-0.4270860\*(4.05)-0.4509639\*(7.27)-0.2423079\*(45.66)-0.3029068\*(34.92)

???

# 第一個主成分的估計值heptathlon\_pca$x[,1]

 實際7項運動的大會計分heptathlon$score

 

求兩者相關係數

cor(heptathlon$score,heptathlon\_pca$x[,1])



高度負的線性相關

# 摘要主成分分析結果(含每個主成分的標準差, 可解釋樣本變異比例與累積可解釋變異比例)

summary(heptathlon\_pca)



# 第一個主成分可解釋61.77%總樣本變異

# 第二個主成分可解釋12.84%總樣本變異

# 兩個主成分累積可解釋74%的樣本變異

# scree plot 陡坡圖(畫出每個主成分可解釋樣本變異比例)

plot(heptathlon\_pca, main = "Scree Plot")



# Score vs C1

# 第一個主成分的估計值heptathlon\_pca$x[,1]

 實際7項運動的大會計分heptathlon$score

# 畫兩者的散佈圖(x軸大會計分, y軸第一個主成分的估計值)

plot(heptathlon$score, heptathlon\_pca$x[,1],

 pch=16, cex=1.5,

 xlab="score",

 ylab="C1")

title(main="Scatter plot")



# 第二個主成分的負荷量(係數)

heptathlon\_pca$rotation[,2]

# 第三個主成分的負荷量(係數)

heptathlon\_pca$rotation[,3]



# 第二個主成分在跳高與800公尺的權重較大,

 可以命名為??

# 第三個主成分在標槍的權重較大,

 可以命名為??

# Biplot: Plot C1 and C2

# a generalization of the simple two-variable scatterplot

biplot(heptathlon\_pca, col=c("gray","black"))



biplot(heptathlon\_pca, col=c("blue","red"))



cbind(heptathlon\_pca$rotation[,1],heptathlon\_pca$rotation[,2])



plot(heptathlon\_pca$rotation[,1],heptathlon\_pca$rotation[,2],

 xlim=c(-1,1), ylim=c(-1, 1))



cbind(heptathlon\_pca$x[,1],heptathlon\_pca$x[,2])



plot(heptathlon\_pca$x[,1],heptathlon\_pca$x[,2])



plot(heptathlon\_pca$x[,1],heptathlon\_pca$x[,2])

text(heptathlon\_pca$x[,1],heptathlon\_pca$x[,2], labels=rownames(heptathlon))

