Homework7

Yi Chen

3/20/2020

## Homework 7

Read data and name it as trackwoman.

trackwoman <- read.table("TrackWomen.dat",sep = "\t",header = T)  
head(trackwoman,5)

## Country X100m X200m X400m X800m X1500m X3000m Marathon  
## 1 ARG 11.57 22.94 52.50 2.05 4.25 9.19 150.32  
## 2 AUS 11.12 22.23 48.63 1.98 4.02 8.63 143.51  
## 3 AUT 11.15 22.70 50.62 1.94 4.05 8.78 154.35  
## 4 BEL 11.14 22.48 51.45 1.97 4.08 8.82 143.05  
## 5 BER 11.46 23.05 53.30 2.07 4.29 9.81 174.18

#### Question 1

# correlation matrix  
R <- cor(trackwoman[,2:ncol(trackwoman)])  
round(R,2)

## X100m X200m X400m X800m X1500m X3000m Marathon  
## X100m 1.00 0.94 0.87 0.81 0.78 0.73 0.67  
## X200m 0.94 1.00 0.91 0.82 0.80 0.73 0.68  
## X400m 0.87 0.91 1.00 0.81 0.72 0.67 0.68  
## X800m 0.81 0.82 0.81 1.00 0.91 0.87 0.85  
## X1500m 0.78 0.80 0.72 0.91 1.00 0.97 0.79  
## X3000m 0.73 0.73 0.67 0.87 0.97 1.00 0.80  
## Marathon 0.67 0.68 0.68 0.85 0.79 0.80 1.00

# eigenvalues  
eigenvalue <- eigen(R)$values  
round(eigenvalue,3)

## [1] 5.808 0.629 0.279 0.125 0.091 0.055 0.014

# eigenvector  
eigenvector <- eigen(R)$vector  
rownames(eigenvector) = colnames(trackwoman[,2:ncol(trackwoman)])  
round(eigenvector,2)

## [,1] [,2] [,3] [,4] [,5] [,6] [,7]  
## X100m -0.38 -0.41 -0.14 0.59 -0.17 -0.54 0.09  
## X200m -0.38 -0.41 -0.10 0.19 0.09 0.74 -0.27  
## X400m -0.37 -0.46 0.24 -0.65 0.33 -0.24 0.13  
## X800m -0.39 0.16 0.15 -0.30 -0.82 0.02 -0.20  
## X1500m -0.39 0.31 -0.42 -0.07 0.03 0.19 0.73  
## X3000m -0.38 0.42 -0.41 -0.08 0.35 -0.24 -0.57  
## Marathon -0.36 0.39 0.74 0.32 0.25 0.05 0.08

#### Question 2

# first two principal components  
(PC2 <- eigenvector[,1:2])

## [,1] [,2]  
## X100m -0.3777657 -0.4071756  
## X200m -0.3832103 -0.4136291  
## X400m -0.3680361 -0.4593531  
## X800m -0.3947810 0.1612459  
## X1500m -0.3892610 0.3090877  
## X3000m -0.3760945 0.4231899  
## Marathon -0.3552031 0.3892153

# the cumulative percentage  
sum(eigen(R)$values[1:2])/sum(eigen(R)$values)

## [1] 0.919474

#### Question 3

For the first principal component, the the level of coefficienct are similar for all variables. The variable that contributes most is X800m.

For the second principal component, the variable has biggest postive effect is X3000m. Meanwhile the variable has the biggest negative effecgt is X400m.

#### Question 4

loading\_1 <-as.matrix(trackwoman[,2:ncol(trackwoman)]) %\*% as.matrix(PC2[,1])  
rownames(loading\_1) <- trackwoman$Country  
rownames(loading\_1)[order(loading\_1[,1],decreasing = T)]

## [1] "GBR" "CHN" "GER" "USA" "RUS" "KEN" "AUS" "ROM"   
## [9] "POL" "CZE" "JPN" "IRL" "MEX" "NOR" "BEL" "ITA"   
## [17] "NED" "FRA" "POR" "ESP" "SUI" "CAN" "FIN" "BRA"   
## [25] "NZL" "HUN" "KOR, S" "SWE" "DEN" "GRE" "ARG" "TUR"   
## [33] "AUT" "KOR, N" "COL" "LUX" "CHI" "IND" "ISR" "INA"   
## [41] "MYA" "TPE" "SIN" "THA" "CRC" "PHI" "DOM" "MAS"   
## [49] "MRI" "BER" "GUA" "SAM" "PNG" "COK"

Broadly, the rank match the my intuitive notion of athletic excellence. For example the countries like GBR, CHN, GER, and USA are among the top. Countries like COK, PNG, SAM, and GUA are among the buttom.

#### Question 5

Convert the variable into the speed using KM.

trackwoman$X100m <- 100/trackwoman$X100m  
trackwoman$X200m <- 200/trackwoman$X200m  
trackwoman$X400m <- 400/trackwoman$X400m  
trackwoman$X800m <- 800/(trackwoman$X800m\*60)  
trackwoman$X1500m <- 1500/(trackwoman$X1500m\*60)  
trackwoman$X3000m <- 3000/(trackwoman$X3000m\*60)  
trackwoman$Marathon <- 42195/(trackwoman$Marathon\*60)  
head(trackwoman)

## Country X100m X200m X400m X800m X1500m X3000m Marathon  
## 1 ARG 8.643042 8.718396 7.619048 6.504065 5.882353 5.440696 4.678353  
## 2 AUS 8.992806 8.996851 8.225375 6.734007 6.218905 5.793743 4.900355  
## 3 AUT 8.968610 8.810573 7.902015 6.872852 6.172840 5.694761 4.556203  
## 4 BEL 8.976661 8.896797 7.774538 6.768190 6.127451 5.668934 4.916113  
## 5 BER 8.726003 8.676790 7.504690 6.441224 5.827506 5.096840 4.037490  
## 6 BRA 8.952551 8.849558 7.902015 6.768190 5.995204 5.530973 4.770708

#### Question 6

R <- cov(trackwoman[,2:ncol(trackwoman)])  
round(R,2)

## X100m X200m X400m X800m X1500m X3000m Marathon  
## X100m 0.09 0.10 0.10 0.07 0.08 0.09 0.08  
## X200m 0.10 0.11 0.11 0.07 0.10 0.11 0.09  
## X400m 0.10 0.11 0.14 0.08 0.10 0.11 0.10  
## X800m 0.07 0.07 0.08 0.07 0.09 0.10 0.09  
## X1500m 0.08 0.10 0.10 0.09 0.12 0.14 0.12  
## X3000m 0.09 0.11 0.11 0.10 0.14 0.18 0.15  
## Marathon 0.08 0.09 0.10 0.09 0.12 0.15 0.17

# eigenvalues  
eigenvalue <- eigen(R)$values  
round(eigenvalue,3)

## [1] 0.732 0.086 0.033 0.015 0.009 0.006 0.002

# eigenvector  
eigenvector <- eigen(R)$vector  
rownames(eigenvector) = colnames(trackwoman[,2:ncol(trackwoman)])  
round(eigenvector,2)

## [,1] [,2] [,3] [,4] [,5] [,6] [,7]  
## X100m -0.31 -0.38 -0.10 0.58 0.05 0.62 0.14  
## X200m -0.36 -0.43 -0.09 0.32 0.03 -0.69 -0.31  
## X400m -0.38 -0.52 0.27 -0.67 0.19 0.12 0.13  
## X800m -0.30 0.05 0.05 -0.13 -0.89 0.14 -0.26  
## X1500m -0.39 0.21 -0.43 -0.06 -0.13 -0.24 0.73  
## X3000m -0.46 0.40 -0.43 -0.18 0.36 0.20 -0.50  
## Marathon -0.42 0.44 0.73 0.24 0.14 -0.08 0.10

#### Question 7

# first two principal components  
(PC2 <- eigenvector[,1:2])

## [,1] [,2]  
## X100m -0.3102442 -0.37596510  
## X200m -0.3573948 -0.43376925  
## X400m -0.3787367 -0.51873227  
## X800m -0.2993405 0.05313551  
## X1500m -0.3912131 0.21084397  
## X3000m -0.4595909 0.39557338  
## Marathon -0.4227291 0.44458346

# the cumulative percentage  
sum(eigen(R)$values[1:2])/sum(eigen(R)$values)

## [1] 0.9259427

#### Question 8

Compared with the first PCA analysis, the proportation of varianece that explained by the first two principal components is higher.

For the first principal component,the differences among different variables’ are similat. The biggest contribution comes from X3000m and marathon, which rely more one the long distance running sport.

For the second principal component, the sports with short distance (smaler or equal to 400m) is nigative, the longer distane the positive effect is more clear. This component is defined based on the distance.

#### Question 4

loading\_1 <-as.matrix(trackwoman[,2:ncol(trackwoman)]) %\*% as.matrix(PC2[,1])  
rownames(loading\_1) <- trackwoman$Country  
rownames(loading\_1)[order(loading\_1[,1],decreasing = T)]

## [1] "SAM" "COK" "PNG" "GUA" "SIN" "DOM" "CRC" "PHI"   
## [9] "MRI" "BER" "INA" "THA" "MAS" "LUX" "KOR, N" "TPE"   
## [17] "CHI" "MYA" "ISR" "KOR, S" "ARG" "COL" "IND" "DEN"   
## [25] "JPN" "SWE" "NZL" "BRA" "NOR" "HUN" "TUR" "GRE"   
## [33] "AUT" "MEX" "SUI" "BEL" "FIN" "KEN" "POR" "IRL"   
## [41] "NED" "ITA" "CAN" "ESP" "AUS" "CZE" "POL" "ROM"   
## [49] "FRA" "GBR" "GER" "RUS" "CHN" "USA"

I need to note that the direction of the loading effect is different from the last PCA analysis. In this time the countries with better performance in running are at the bottom, while the countries with worse performance are t the top. In general, the result is similar.