HW_PCA

공소연

2022-09-24

Q1: Load “decathlon2” dataset and create a new dataset excluding the “Rank” and “Competition” variables.

```
library(factoextra)
```

```
## 필요한 패키지를 로딩중입니다: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa
```

```
data("decathlon2")
View(decathlon2)

data <- decathlon2[,-c(11, 13)]
```

Q2: Use the “Points” variable as the dependent variable and create the independent variable(x) and dependent

```
y <- data.frame(data[,c(11)])
x <- data[,c(1:10)]
```

Q3: Conduct a principal component analysis using independent variable set and check the importance of components.

```
pcs <- prcomp(na.omit(x), scale. = T)
summary(pcs)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6
PC7
## Standard deviation     1.936 1.3210 1.2320 1.0160 0.78603 0.65444
0.57089
## Proportion of Variance 0.375 0.1745 0.1518 0.1032 0.06178 0.04283
0.03259
## Cumulative Proportion  0.375 0.5495 0.7013 0.8045 0.86630 0.90913
0.94172
##                            PC8     PC9    PC10
## Standard deviation     0.52857 0.43716 0.33511
## Proportion of Variance 0.02794 0.01911 0.01123
## Cumulative Proportion  0.96966 0.98877 1.00000
```

```
pcs$rotation
```

```
##                       PC1          PC2          PC3          PC4
PC5
```

```
## X100m        -0.42290657  0.2594748 -0.081870461  0.09974877 -
0.2796419
## Long.jump     0.39189495 -0.2887806  0.005082180 -0.18250903
0.3355025
## Shot.put      0.36926619  0.2135552 -0.384621732  0.03553644 -
0.3544877
## High.jump     0.31422571  0.4627797 -0.003738604  0.07012348
0.3824125
## X400m        -0.33248297  0.1123521 -0.418635317  0.26554389
0.2534755
## X110m.hurdle -0.36995919  0.2252392 -0.338027983 -0.15726889
0.2048540
## Discus        0.37020078  0.1547241 -0.219417086  0.39137188 -
0.4319091
## Pole.vault   -0.11433982 -0.5583051 -0.327177839 -0.24759476 -
0.3340758
## Javeline      0.18341259  0.0745854 -0.564474643 -0.47792535
0.1697426
## X1500m        0.03599937 -0.4300522 -0.286328973  0.64220377
0.3227349
##                      PC6         PC7         PC8         PC9
PC10
## X100m         0.16023494 -0.03227949  0.35266427 -0.71190625
0.03272397
## Long.jump     0.07384658  0.24902853  0.72986071 -0.12801382
0.02395904
## Shot.put      0.32207320  0.23059438 -0.01767069  0.07184807 -
0.61708920
## High.jump     0.52738027  0.03992994 -0.25003572 -0.14583529
0.41523052
## X400m        -0.23884715  0.69014364 -0.01543618  0.13706918
0.12016951
## X110m.hurdle  0.26249611 -0.42797378  0.36415520  0.49550598 -
0.03514180
## Discus       -0.28217086 -0.18416631  0.26865454  0.18621144
0.48037792
## Pole.vault    0.43606610  0.12654370 -0.16086549  0.02983660
0.40290423
## Javeline     -0.42368592 -0.23324548 -0.19922452 -0.33300936
0.02100398
## X1500m        0.10850981 -0.34406521 -0.09752169 -0.19899138 -
0.18954698
```

Q4: Choose some components to conduct a regression analysis to predict the dependent variable. How many components did you choose? Explain.

PC1부터 PC6까지 6개를 선택했다. PC6에서 Cumulative Proportion이 약 91%인데, 이는 PC6까지 주성분을 선택할시 전체 변동성의 91%를 설명해준다는 뜻이다. 즉 이 자료를 91% 정도까지 설명해줄 수 있다는 것이므로 6개를 선택했다.