Homework8

Yi Chen

4/7/2020

## Homework 8

## read the table   
trackwomen <- read.table("TrackWomen.dat", header = TRUE,sep = "\t")  
head(trackwomen)

## 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  
## 6 BRA 11.17 22.60 50.62 1.97 4.17 9.04 147.41

#### question a

## create the correlation matrix  
library(psych)  
(trackwomen.fa = principal(trackwomen[2:ncol(trackwomen)],   
 nfactors = 2,   
 rotate = 'none',  
 scores = TRUE,  
 covar = FALSE))

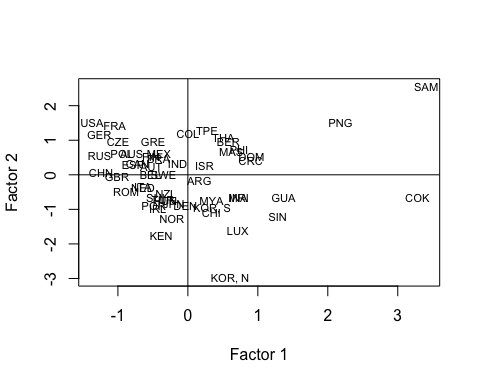
## Principal Components Analysis  
## Call: principal(r = trackwomen[2:ncol(trackwomen)], nfactors = 2, rotate = "none",   
## covar = FALSE, scores = TRUE)  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## PC1 PC2 h2 u2 com  
## X100m 0.91 -0.32 0.93 0.067 1.2  
## X200m 0.92 -0.33 0.96 0.040 1.2  
## X400m 0.89 -0.36 0.92 0.081 1.3  
## X800m 0.95 0.13 0.92 0.079 1.0  
## X1500m 0.94 0.25 0.94 0.060 1.1  
## X3000m 0.91 0.34 0.93 0.066 1.3  
## Marathon 0.86 0.31 0.83 0.172 1.3  
##   
## PC1 PC2  
## SS loadings 5.81 0.63  
## Proportion Var 0.83 0.09  
## Cumulative Var 0.83 0.92  
## Proportion Explained 0.90 0.10  
## Cumulative Proportion 0.90 1.00  
##   
## Mean item complexity = 1.2  
## Test of the hypothesis that 2 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0.03   
## with the empirical chi square 2.63 with prob < 0.96   
##   
## Fit based upon off diagonal values = 1

The second factor one has high loading on the short-disctance sports, which can be interpreated as explosiveness (except x800m).

While the first factor has high loading on the long-distance sport, which can be interpreated as endurance.

#### question b

plot(trackwomen.fa$scores, xlab = "Factor 1", ylab = 'Factor 2',type = 'n')  
text(trackwomen.fa$scores,labels = trackwomen$Country,cex=.7)  
abline(h = 0, v = 0)



As we can see, SAM, COK, and PNG are three outliers in the first factor.

While KOR, N is the outlier in the second factor.

#### question c

(trackwomen.fa.var = principal(trackwomen[2:ncol(trackwomen)], nfactors = 2, rotate = 'varimax',scores = TRUE,covar = FALSE))

## Principal Components Analysis  
## Call: principal(r = trackwomen[2:ncol(trackwomen)], nfactors = 2, rotate = "varimax",   
## covar = FALSE, scores = TRUE)  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## RC1 RC2 h2 u2 com  
## X100m 0.43 0.86 0.93 0.067 1.5  
## X200m 0.44 0.88 0.96 0.040 1.5  
## X400m 0.39 0.88 0.92 0.081 1.4  
## X800m 0.77 0.57 0.92 0.079 1.8  
## X1500m 0.85 0.48 0.94 0.060 1.6  
## X3000m 0.89 0.39 0.93 0.066 1.4  
## Marathon 0.83 0.37 0.83 0.172 1.4  
##   
## RC1 RC2  
## SS loadings 3.31 3.13  
## Proportion Var 0.47 0.45  
## Cumulative Var 0.47 0.92  
## Proportion Explained 0.51 0.49  
## Cumulative Proportion 0.51 1.00  
##   
## Mean item complexity = 1.5  
## Test of the hypothesis that 2 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0.03   
## with the empirical chi square 2.63 with prob < 0.96   
##   
## Fit based upon off diagonal values = 1

The second factor one has high loading on the short-disctance sports, which can be interpreated as explosiveness.

While the first factor has high loading on the long-distance sport, which can be interpreated as endurance.

This time all the loadings are positive and the high and low value are more clearly distinguished for the sport below the 400m and above 400m.

#### question d

trackwomen.fa.mle = factanal(covmat = cor(trackwomen[2:ncol(trackwomen)]),   
 factors = 3,   
 rotation = "none")  
trackwomen.fa.mle$loadings

##   
## Loadings:  
## Factor1 Factor2 Factor3  
## X100m 0.905 -0.273   
## X200m 0.941 -0.331   
## X400m 0.858 -0.316 0.177   
## X800m 0.920 0.120 0.304   
## X1500m 0.954 0.291   
## X3000m 0.908 0.368   
## Marathon 0.789 0.173 0.350   
##   
## Factor1 Factor2 Factor3  
## SS loadings 5.641 0.549 0.249  
## Proportion Var 0.806 0.078 0.036  
## Cumulative Var 0.806 0.884 0.920

It becomes harder to interperated the three factors. The first factor has high loading almost on all the sports. The second factor has high loading value on the long-distance sprots. The thrid factor onle depend on the x400m, x800m, and marathon.

#### question e

trackwomen.fa.mle = factanal(covmat = cor(trackwomen[2:ncol(trackwomen)]),  
 factors = 3,  
 rotation = "varimax")  
trackwomen.fa.mle$loadings

##   
## Loadings:  
## Factor1 Factor2 Factor3  
## X100m 0.815 0.413 0.245   
## X200m 0.886 0.410 0.203   
## X400m 0.797 0.311 0.367   
## X800m 0.512 0.617 0.556   
## X1500m 0.449 0.849 0.270   
## X3000m 0.361 0.866 0.280   
## Marathon 0.380 0.553 0.571   
##   
## Factor1 Factor2 Factor3  
## SS loadings 2.824 2.593 1.022  
## Proportion Var 0.403 0.370 0.146  
## Cumulative Var 0.403 0.774 0.920

The first factor has high loading on the short-distance sport. The second factor has high loading value on the long-distance sprots. The thrid factor onle depend on the x400m and marathon.

All factors have a negative loading on the x800m.

#### question f

trackwomen.fa.mle = factanal(covmat = cor(trackwomen[2:ncol(trackwomen)]),  
 factors = 3,   
 rotation = "none")  
trackwomen.fa.mle

##   
## Call:  
## factanal(factors = 3, covmat = cor(trackwomen[2:ncol(trackwomen)]), rotation = "none")  
##   
## Uniquenesses:  
## X100m X200m X400m X800m X1500m X3000m Marathon   
## 0.106 0.005 0.133 0.047 0.005 0.041 0.225   
##   
## Loadings:  
## Factor1 Factor2 Factor3  
## X100m 0.905 -0.273   
## X200m 0.941 -0.331   
## X400m 0.858 -0.316 0.177   
## X800m 0.920 0.120 0.304   
## X1500m 0.954 0.291   
## X3000m 0.908 0.368   
## Marathon 0.789 0.173 0.350   
##   
## Factor1 Factor2 Factor3  
## SS loadings 5.641 0.549 0.249  
## Proportion Var 0.806 0.078 0.036  
## Cumulative Var 0.806 0.884 0.920  
##   
## The degrees of freedom for the model is 3 and the fit was 0.2033

The chi-square test statistics is 9.73 with 3 degree of freedom, the p-value is 0.2221. The null hypothesis, H0, is that the number of factors in the model, in our example 2 factors, is sufficient to capture the full dimensionality of the data set. Conventionally, we reject H0 if the p-value is less than 0.05. Such a result indicates that the number of factors is too small. In contrast, we do not reject H0 if the p-value exceeds 0.05. In this case, we could conclude that there is a significant evidence of model fit and the number of factor is enough in the model.

#### question g

trackwomen$X100m <- 100/trackwomen$X100m  
trackwomen$X200m <- 200/trackwomen$X200m  
trackwomen$X400m <- 400/trackwomen$X400m  
trackwomen$X800m <- 800/(trackwomen$X800m\*60)  
trackwomen$X1500m <- 1500/(trackwomen$X1500m\*60)  
trackwomen$X3000m <- 3000/(trackwomen$X3000m\*60)  
trackwomen$Marathon <- 42195/(trackwomen$Marathon\*60)  
  
## repeat the d  
(trackwomen.fa.mle = factanal(covmat = cov(trackwomen[2:ncol(trackwomen)]),   
 factors = 3,   
 rotation = "none"))

##   
## Call:  
## factanal(factors = 3, covmat = cov(trackwomen[2:ncol(trackwomen)]), rotation = "none")  
##   
## Uniquenesses:  
## X100m X200m X400m X800m X1500m X3000m Marathon   
## 0.105 0.016 0.142 0.113 0.005 0.036 0.091   
##   
## Loadings:  
## Factor1 Factor2 Factor3  
## X100m 0.843 0.428   
## X200m 0.876 0.466   
## X400m 0.803 0.441 0.136   
## X800m 0.925 0.178   
## X1500m 0.988 -0.130   
## X3000m 0.958 -0.205   
## Marathon 0.836 -0.109 0.445   
##   
## Factor1 Factor2 Factor3  
## SS loadings 5.572 0.666 0.253  
## Proportion Var 0.796 0.095 0.036  
## Cumulative Var 0.796 0.891 0.927  
##   
## The degrees of freedom for the model is 3 and the fit was 0.209

## repeat the e  
(trackwomen.fa.mle = factanal(covmat = cov(trackwomen[2:ncol(trackwomen)]),   
 factors = 3,   
 rotation = "varimax"))

##   
## Call:  
## factanal(factors = 3, covmat = cov(trackwomen[2:ncol(trackwomen)]), rotation = "varimax")  
##   
## Uniquenesses:  
## X100m X200m X400m X800m X1500m X3000m Marathon   
## 0.105 0.016 0.142 0.113 0.005 0.036 0.091   
##   
## Loadings:  
## Factor1 Factor2 Factor3  
## X100m 0.837 0.400 0.184   
## X200m 0.890 0.408 0.159   
## X400m 0.810 0.338 0.295   
## X800m 0.547 0.671 0.372   
## X1500m 0.480 0.855 0.181   
## X3000m 0.392 0.859 0.270   
## Marathon 0.355 0.635 0.617   
##   
## Factor1 Factor2 Factor3  
## SS loadings 2.958 2.763 0.771  
## Proportion Var 0.423 0.395 0.110  
## Cumulative Var 0.423 0.817 0.927  
##   
## The degrees of freedom for the model is 3 and the fit was 0.209

#### question h

I personlly perfer the model after the transformation, which make the data in the same scale as speed and reduce the error from the scale.

I perfer the factor analysis based on the covariance matrix with varimax rotation since the loading is easier to interpreate.

The first factor has the high loading value on the short-distance sports, the seconf factor has the high loading value on the long-distance sports, and the thrid factor has the high loading value on x400, x800m, and marathon.