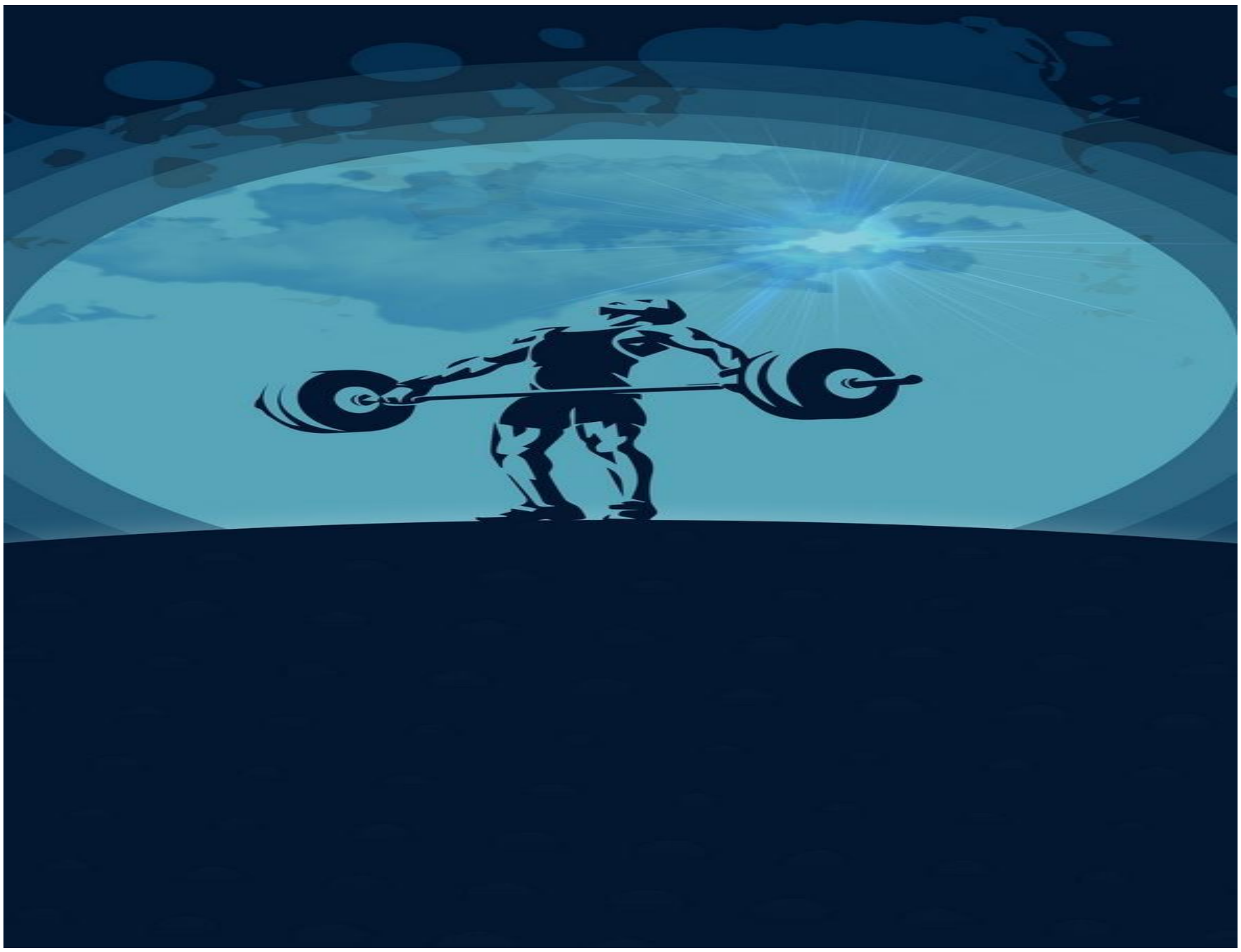
Factor Analysis Report for "Body Fat Prediction" Dataset



S/16/844

W.Kavinda Uthsuka

1. Introduction

- Factor analysis is a statistical method used to search for some unobserved variables called factors from observed variables called factors.
- I use 'Body Fat Prediction' dataset to perform the factor analysis.
- Test the hypothesis that the selected factors are sufficient.
- Main objective is reducing the dimensions into small number of dimensions and convert them to interpretable way.

2. Methodology

I use 'Body Fat Prediction' dataset which estimates body fat and various body circumference measurements for 252 men. The dataset describes these based on 15 variables;

- Density – Density determined from underwater weighing
- BodyFat – Percent body fat from Siri's (1965) equation
- Age – Age from years
- Weight – Weight from lbs
- Height – Height from inches
- Neck – Neck circumference (cm)
- Chest – Chest circumference (cm)
- Abdomen – Abdomen 2 circumference (cm)
- Hip – Hip circumference (cm)
- Thigh – Thigh circumference (cm)
- Knee – Knee circumference (cm)
- Ankle – Ankle circumference (cm)
- Biceps – Biceps (extended) circumference (cm)
- Forearm – Forearm circumference (cm)
- Wrist – Wrist circumference (cm)

Use Methods;

- Exploratory Factor Analysis
- Confirmatory Factor Analysis

3. Results and Discussion

- From standardized dataset I can put same weight for all the variables.
- KMO test output

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = df_st)
## Overall MSA =  0.91
## MSA for each item =
## Density BodyFat     Age  Weight  Height    Neck   Chest Abdomen     H
ip   Thigh
##    0.81    0.82    0.43    0.89    0.62    0.95    0.93    0.94    0.
92    0.93
##    Knee   Ankle  Biceps Forearm   Wrist
##    0.95    0.94    0.96    0.94    0.93
```

Overall MSA = 0.91 (>0.9). Therefore we can conclude that this dataset is ve ry good to use for factor analysis.

- Eigen values

```
##  [1] 8.92101309 1.90252302 1.07304831 0.70061268 0.65035002 0.50098894
##  [7] 0.30551840 0.26159115 0.21761584 0.18093818 0.13151081 0.07698811
## [13] 0.04247634 0.02344280 0.01138231
```

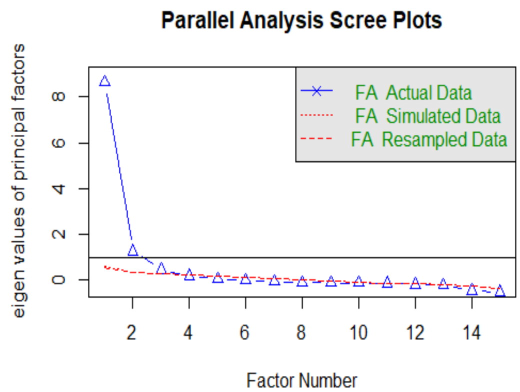
From Kaiser (1961) method, we take factors which are eigenvalue should be at least greater than one. For this dataset we can get first three eigenvalues. Therefore our factor model has only three factors.

- Proportion of variance

```
##  [1] 0.5947342059 0.1268348679 0.0715365538 0.0467075121 0.0433566682
##  [6] 0.0333992628 0.0203678934 0.0174394100 0.0145077229 0.0120625455
## [11] 0.0087673876 0.0051325405 0.0028317558 0.0015628531 0.0007588205
```

Cumulative proportion variance explained by first three factors = 0.7931. Therefore we can conclude that factor model explains 79.31% of total variance and this is a good model.

- Scree Plot and Parallel Analysis Scree Plots




```
## Parallel analysis suggests that the number of factors = 3  and the
number of components =  NA
```

- Hypothesis Testing

H null: Test the hypothesis that are three factors are sufficient

H alternative: More factors are needed

```
## The harmonic number of observations is 252 with the empirical chi
square 44.09  with prob <  0.97
```

Model probability value (0.97) is greater than 0.05. Therefore we can conclude that 3 factor model is sufficient at 5% significance level.

- PC Factor loadings

```
##                PA1         PA2          PA3
## Density -0.70854971  0.62239844  0.163163210
## BodyFat  0.72480259 -0.62652415 -0.145219069
## Age      0.07778733 -0.53520611  0.575376427
## Weight   0.97614179  0.14709196 -0.025356758
## Height   0.21293607  0.40362645  0.118785451
## Neck     0.83571360  0.12271629  0.201821702
## Chest    0.90894698 -0.11904491  0.073961893
## Abdomen  0.92688570 -0.26496640  0.008287817
```

```
## Hip      0.92333486  0.06677004 -0.171221701
## Thigh    0.87218940  0.16211045 -0.319143397
## Knee     0.84334549  0.16869442 -0.006405298
## Ankle    0.58718447  0.27924240  0.023618174
## Biceps   0.81439753  0.16242990 -0.021656458
## Forearm  0.65205054  0.21553182  0.045156073
## Wrist    0.75099967  0.23482863  0.463603507
```

In factor 1 (PA1) , there is contrast between 'Density' and other variables. In factor 2 (PA2), there is a contrast between 'BodyFat' , 'Age', 'Chest' , 'Abdomen' and other variables. In factor 3 (PA3), there is a contrast between 'BodyFat', 'Weight', 'Hip', 'Thigh', 'Knee', 'Biceps' and other variables. Factor loadings are not giving clear conclusion about the model. Therefore we have to rotate them.

- Factor analyze using Maximum Likelihood Method

```
##                         ML1  ML2  ML3
## SS loadings            7.28 2.91 0.75
## Proportion Var         0.49 0.19 0.05
## Cumulative Var         0.49 0.68 0.73
## Proportion Explained   0.67 0.27 0.07
## Cumulative Proportion  0.67 0.93 1.00
```

ML method explains 73% of total sample variance of the dataset. But PC method can explains 79.13% total sample variance of the dataset. Therefore we can conclude that PC method is the best method to do factor analysis for this dataset.

4. Conclusion and Recommendation

- According to the analysis we can get three factors. I proved from empirical chi squared test.
- 3 factor model explains greater than 79% total variance of the dataset. So, this is enough to interpret the idea.
- After rotating the factor loadings from "varimax" method;

```
##                PA1        PA2         PA3
## Density -0.15268021 -0.9304430 -0.16434487
## BodyFat  0.16843824  0.9367094  0.18210963
## Age     -0.06146188  0.1915626  0.76359720
## Weight   0.82908505  0.5290307 -0.08874851
## Height   0.42513334 -0.1822244 -0.09177141
## Neck     0.77624406  0.3702176  0.12079987
```

```
## Chest    0.65961267  0.6281224  0.12727957
## Abdomen  0.57240894  0.7627795  0.14109731
## Hip      0.69911167  0.6050916 -0.17728341
## Thigh    0.66735130  0.5645864 -0.35317284
## Knee     0.74413567  0.4231263 -0.08340699
## Ankle    0.61652792  0.1748951 -0.11235999
## Biceps   0.71360838  0.4147962 -0.09382434
## Forearm  0.63789854  0.2507627 -0.06218101
## Wrist    0.85286301  0.1415743  0.29437155
```

The rotated loadings indicate that the variables West, Neck, Chest, Hip, Thigh, Knee, Ankle, Biceps, Wrist load highly on the first factor (PA1). Density, BodyFat, Abdomen load highly on second factor (PA2) and Age describes from third factor (PA3). We might call factor 1 as “General factor”, factor 2 as “Secondary factor” and factor 3 as “Tertiary factor”. We can name the factors based on number of variables highly load by each factor.

- Communalities

```
## Density                                         0.9160447
## BodyFat                                         0.9389599
## Age                                             0.6235545
## Weight                                          0.9751318
## Height                                          0.2223661
## Neck                                            0.7542085
## Chest                                           0.8458267
## Abdomen                                         0.9293930
## Hip                                             0.8863224
## Thigh                                           0.8888467
## Knee                                            0.7397304
## Ankle                                           0.4233197
## Biceps                                          0.6900958
## Forearm                                         0.4736629
## Wrist                                           0.8340732
```

The model explains Density, Bodyfat, Weight, Abdomen, Hip, Thigh the best and is not bad for other variables such as Age, Neck, Knee, Chest, Biceps, Wrist. However, for other variables such as Height, Ankle, Forearm the model doesn’t do a good job, explaining only under the half of the variation.

5. References

References

Dang, A. (n.d.). *Exploratory Factor Analysis in R.*

Pramoditha, R. (n.d.). *Factor Analysis on "Women Track Records" Data with R and Python.*

6. Appendices

- Part of the dataset - https://www.kaggle.com/fedesoriano/body-fat-prediction-dataset

Density	BodyFat	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
1.0708	12.3	23	154.25	67.75	36.2	93.1	85.2	94.5	59	37.3	21.9	32	27.4	17.1
1.0853	6.1	22	173.25	72.25	38.5	93.6	83	98.7	58.7	37.3	23.4	30.5	28.9	18.2
1.0414	25.3	22	154	66.25	34	95.8	87.9	99.2	59.6	38.9	24	28.8	25.2	16.6
1.0751	10.4	26	184.75	72.25	37.4	101.8	86.4	101.2	60.1	37.3	22.8	32.4	29.4	18.2
1.034	28.7	24	184.25	71.25	34.4	97.3	100	101.9	63.2	42.2	24	32.2	27.7	17.7
1.0502	20.9	24	210.25	74.75	39	104.5	94.4	107.8	66	42	25.6	35.7	30.6	18.8
1.0549	19.2	26	181	69.75	36.4	105.1	90.7	100.3	58.4	38.3	22.9	31.9	27.8	17.7
1.0704	12.4	25	176	72.5	37.8	99.6	88.5	97.1	60	39.4	23.2	30.5	29	18.8
1.09	4.1	25	191	74	38.1	100.9	82.5	99.9	62.9	38.3	23.8	35.9	31.1	18.2

- Codes

```
library(data.table)

library(factoextra)

library(psych)

library(corrplot)

library(ggplot2)

df <- fread("bodyfat.csv",header = TRUE)
head(df)

df[is.na(df)] <- 0

describe(df)

df_st <- apply(df,2,scale)
df_st

KMO(df_st)

df_st_cov <- cov(df_st)
df_st_cov
```

```r
df_st_cov_eigen <- eigen(df_st_cov)
df_st_cov_eigen$values

df_st_cov_eigen$vectors

PVE <- df_st_cov_eigen$values / sum(df_st_cov_eigen$values)
PVE

scree(df_st)

fa.parallel(df_st,fm="pa",fa="fa")

df_st_fa_pc <- fa(df_st_cov ,nfactors = 3,rotate = "none",n.obs = 252
,covar = TRUE,fm = "pa")
df_st_fa_pc

df_st_fa_pc$loadings

unrotated_pc_loadings <- as.data.frame(unclass(df_st_fa_pc$loadings))
unrotated_pc_loadings

unrotated_pc_com <- as.data.frame(unclass(df_st_fa_pc$communality))
unrotated_pc_com

df_st_fa_ml <- fa(df_st_cov,nfactors = 3,rotate = "none",n.obs = 252
, covar = TRUE, fm = 'ml')
df_st_fa_ml

df_st_fa_ml$loadings

unrotated_ml_loadings <- as.data.frame(unclass(df_st_fa_ml$loadings))
unrotated_ml_loadings

unrotated_ml_com <- as.data.frame(unclass(df_st_fa_ml$communality))
unrotated_ml_com

library(GPArotation)
df_st_fa_pc_rotate <- fa(df_st_cov ,nfactors = 3,rotate = "varimax",n
.obs = 252 ,covar = TRUE,fm = 'pa')
df_st_fa_pc_rotate

df_st_fa_pc_rotate$loadings

rotated_pc_loadings <- as.data.frame(unclass(df_st_fa_pc_rotate$loadi
ngs))
rotated_pc_loadings

rotated_pc_com <- as.data.frame(unclass(df_st_fa_pc_rotate$communalit
y))
rotated_pc_com
```