Practice problems for FA with partial answers

1. This problem comes from Chapter 6 of Johnson’s textbook. Below is the description from the book:

The data consist of anthropometric and physical fitness measurements that were taken on 50 applicants to the police department of a major metropolitan city. The variables include:

1. Reaction time in seconds to a visual stimulus (REACT)
2. The applicant’s height in centimeters (HEIGHT)
3. The applicant’s weight in kilograms (WEIGHT)
4. The applicant’s shoulder width in centimeters (SHLDR)
5. The applicant’s pelvic width in centimeters (PELVIC)
6. The applicant’s minimum chest circumference in centimeters (CHEST)
7. The applicant’s thigh skinfold thickness in millimeters (THIGH)
8. The applicant’s resting pulse rate (PULSE)
9. The applicant’s diastolic blood pressure (DIAST)
10. The number of chin-ups the applicant was able to complete (CHNUP)
11. The applicant’s maximum breathing capacity in liters (BREATH)
12. The applicant’s pulse rate after 5 minutes of recovery from treadmill running (RECVR)
13. The applicants maximum treadmill speed (SPEED)
14. The applicant’s treadmill endurance time in minutes (ENDUR)
15. The applicant’s total body fat measurement (FAT)

The data is in the file PoliceApplicant.csv. Using this data, complete the following.

* 1. Use PCA with the correlation matrix to help choose an initial number of common factors.

There are 5 PCs with eigenvalues greater than 1, and they account for 76% of the total variance of 15 variables. For an initial choice, this appears to be a good place to start.

* 1. Using the initial number of common factors from part a), examine the appropriate measures to judge their adequacy.
		1. LRT involving 5 common factors

Part of the output is shown below

Test of the hypothesis that 5 factors are sufficient.

The chi square statistic is 53.8 on 40 degrees of freedom.

The p-value is 0.0712

There is marginal evidence that more factors are needed.

* + 1. 

I stored the residuals in an object named resid5:

> resid5

 REACT HEIGHT WEIGHT SHLDR PELVIC CHEST

REACT 0.0000 0.0003 0.0005 0.0355 -0.0520 -0.0026

HEIGHT 0.0003 0.0000 -0.0014 -0.0101 0.0112 0.0020

WEIGHT 0.0005 -0.0014 0.0000 0.0081 0.0046 -0.0011

SHLDR 0.0355 -0.0101 0.0081 0.0000 0.0241 0.0054

PELVIC -0.0520 0.0112 0.0046 0.0241 0.0000 -0.0304

CHEST -0.0026 0.0020 -0.0011 0.0054 -0.0304 0.0000

THIGH -0.0008 -0.0010 0.0000 0.0127 -0.0018 -0.0020

PULSE 0.0294 -0.0083 0.0058 0.0612 -0.0544 -0.0142

DIAST 0.0333 -0.0192 0.0053 0.0237 0.2750 -0.0188

CHNUP 0.0161 0.0091 -0.0082 -0.0344 0.1140 0.0177

BREATH -0.0191 0.0151 -0.0032 -0.0638 -0.0061 0.0119

RECVR -0.0002 0.0001 0.0000 0.0001 -0.0005 0.0000

SPEED -0.0004 0.0107 -0.0045 0.0327 -0.1014 -0.0051

ENDUR -0.0342 0.0284 0.0044 -0.0150 -0.0174 -0.0046

FAT 0.0028 0.0038 -0.0005 -0.0296 0.0134 0.0055

 THIGH PULSE DIAST CHNUP BREATH RECVR

REACT -0.0008 0.0294 0.0333 0.0161 -0.0191 -2e-04

HEIGHT -0.0010 -0.0083 -0.0192 0.0091 0.0151 1e-04

WEIGHT 0.0000 0.0058 0.0053 -0.0082 -0.0032 0e+00

SHLDR 0.0127 0.0612 0.0237 -0.0344 -0.0638 1e-04

PELVIC -0.0018 -0.0544 0.2750 0.1140 -0.0061 -5e-04

CHEST -0.0020 -0.0142 -0.0188 0.0177 0.0119 0e+00

THIGH 0.0000 0.0001 0.0124 -0.0037 -0.0018 -1e-04

PULSE 0.0001 0.0000 0.0802 0.1046 -0.0504 0e+00

DIAST 0.0124 0.0802 0.0000 0.0929 -0.1133 -1e-04

CHNUP -0.0037 0.1046 0.0929 0.0000 -0.1170 4e-04

BREATH -0.0018 -0.0504 -0.1133 -0.1170 0.0000 2e-04

RECVR -0.0001 0.0000 -0.0001 0.0004 0.0002 0e+00

SPEED -0.0117 0.0038 -0.0881 0.1522 -0.0299 -1e-04

ENDUR 0.0056 -0.0406 0.1297 -0.0714 -0.1400 -2e-04

FAT 0.0001 0.0065 -0.0092 0.0094 -0.0126 2e-04

 SPEED ENDUR FAT

REACT -0.0004 -0.0342 0.0028

HEIGHT 0.0107 0.0284 0.0038

WEIGHT -0.0045 0.0044 -0.0005

SHLDR 0.0327 -0.0150 -0.0296

PELVIC -0.1014 -0.0174 0.0134

CHEST -0.0051 -0.0046 0.0055

THIGH -0.0117 0.0056 0.0001

PULSE 0.0038 -0.0406 0.0065

DIAST -0.0881 0.1297 -0.0092

CHNUP 0.1522 -0.0714 0.0094

BREATH -0.0299 -0.1400 -0.0126

RECVR -0.0001 -0.0002 0.0002

SPEED 0.0000 -0.0951 0.0353

ENDUR -0.0951 0.0000 -0.0191

FAT 0.0353 -0.0191 0.0000

There are a lot of values to examine, so I used the following code to highlight those residuals that may be large in absolute value:

> abs(resid5)>0.1

 REACT HEIGHT WEIGHT SHLDR PELVIC CHEST THIGH PULSE

REACT FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

HEIGHT FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

WEIGHT FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

SHLDR FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

PELVIC FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

CHEST FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

THIGH FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

PULSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

DIAST FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE

CHNUP FALSE FALSE FALSE FALSE TRUE FALSE FALSE TRUE

BREATH FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

RECVR FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

SPEED FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE

ENDUR FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

FAT FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

 DIAST CHNUP BREATH RECVR SPEED ENDUR FAT

REACT FALSE FALSE FALSE FALSE FALSE FALSE FALSE

HEIGHT FALSE FALSE FALSE FALSE FALSE FALSE FALSE

WEIGHT FALSE FALSE FALSE FALSE FALSE FALSE FALSE

SHLDR FALSE FALSE FALSE FALSE FALSE FALSE FALSE

PELVIC TRUE TRUE FALSE FALSE TRUE FALSE FALSE

CHEST FALSE FALSE FALSE FALSE FALSE FALSE FALSE

THIGH FALSE FALSE FALSE FALSE FALSE FALSE FALSE

PULSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE

DIAST FALSE FALSE TRUE FALSE FALSE TRUE FALSE

CHNUP FALSE FALSE TRUE FALSE TRUE FALSE FALSE

BREATH TRUE TRUE FALSE FALSE FALSE TRUE FALSE

RECVR FALSE FALSE FALSE FALSE FALSE FALSE FALSE

SPEED FALSE TRUE FALSE FALSE FALSE FALSE FALSE

ENDUR TRUE FALSE TRUE FALSE FALSE FALSE FALSE

FAT FALSE FALSE FALSE FALSE FALSE FALSE FALSE

> sum(abs(resid5)>0.1)

[1] 18

> sum(abs(resid5)>0.2)

[1] 2

> max(abs(resid5))

[1] 0.275

> colMeans(abs(resid5))

 REACT HEIGHT WEIGHT SHLDR

0.0151466667 0.0080466667 0.0031733333 0.0237600000

 PELVIC CHEST THIGH PULSE

0.0470866667 0.0080866667 0.0035866667 0.0306333333

 DIAST CHNUP BREATH RECVR

0.0600800000 0.0500733333 0.0389600000 0.0001466667

 SPEED ENDUR FAT

0.0380666667 0.0403800000 0.0098666667

> #Number of unique off-diagonal elements

> choose(n = 15, k = 2)

[1] 105

Note that  will always be a symmetric matrix, so there are 105 possible residuals that we need to investigate (diagonal elements of the matrix are always 0 due to the specific variances being added to the common factor part). The largest mean absolute deviation is for the DIAST variable with a value of 0.06. There are 18/2 = 9 residuals that have an absolute value greater than 0.1, and 2/2 = 1 residual that has an absolute value greater than 0.2. Overall, this is not too bad, but it does not hurt to investigate what would happen with 6 common factors.

* + 1. LRT for a different number of common factors

When using 6 common factors, the LRT gives a p-value of 0.17. There are 5 residuals with an absolute value greater than 0.1, and 1 residual with an absolute value greater than 0.2. The maximum possible residual in absolute value is 0.21. Overall, this is an improvement, but may be not enough to persuade me to use 6 common factors.

With respect to using a different number of common factors, below are the LRT results:

 common.fact pvalue

1 1 0.0000

2 2 0.0000

3 3 0.0010

4 4 0.0177

5 5 0.0712

6 6 0.1732

7 7 0.6154

It appears that at least 4 are necessary. Note that with 4 common factors, there are 17 residuals with an absolute value greater than 0.1, and 3 residuals with an absolute value greater than 0.2. The maximum possible residual in absolute value is 0.26.

Overall, I will choose 5 common factors, but other choices are justifiable.

* 1. Using the varimax method, state the FA model for the number of common factors chosen. Interpret the common factors.

Notice the alternative way to include the variables with the x argument.

> mod.fit5v <- factanal(x = set1[,-1], factors = 5, rotation = "varimax")

> print(x = mod.fit5v, cutoff = 0.0)

Call:

factanal(x = set1[, -1], factors = 5, rotation = "varimax")

Uniquenesses:

 REACT HEIGHT WEIGHT SHLDR PELVIC CHEST THIGH PULSE

 0.370 0.109 0.028 0.313 0.485 0.081 0.055 0.621

 DIAST CHNUP BREATH RECVR SPEED ENDUR FAT

 0.870 0.465 0.587 0.005 0.522 0.826 0.058

Loadings:

 Factor1 Factor2 Factor3 Factor4 Factor5

REACT 0.092 0.075 0.060 -0.011 0.782

HEIGHT 0.176 0.888 -0.164 -0.096 0.189

WEIGHT 0.614 0.615 -0.187 0.424 -0.040

SHLDR 0.193 0.747 -0.146 0.100 -0.247

PELVIC 0.238 0.585 -0.272 0.195 -0.066

CHEST 0.488 0.458 -0.112 0.666 -0.121

THIGH 0.957 0.060 0.104 -0.117 0.042

PULSE -0.079 -0.114 0.575 -0.089 0.146

DIAST 0.037 -0.166 0.230 0.166 0.142

CHNUP -0.690 -0.175 -0.028 -0.109 -0.124

BREATH 0.166 0.598 0.082 0.011 0.145

RECVR 0.102 0.059 0.948 -0.127 -0.258

SPEED -0.191 0.166 -0.534 -0.327 -0.147

ENDUR -0.354 -0.198 -0.028 -0.088 0.038

FAT 0.895 0.245 0.017 0.273 0.080

 Factor1 Factor2 Factor3 Factor4 Factor5

SS loadings 3.149 2.843 1.760 0.948 0.905

Proportion Var 0.210 0.190 0.117 0.063 0.060

Cumulative Var 0.210 0.399 0.517 0.580 0.640

Test of the hypothesis that 5 factors are sufficient.

The chi square statistic is 53.8 on 40 degrees of freedom.

The p-value is 0.0712

Part of the model:

z1 = 0.092f1 + 0.075f2 + 0.060f3 – 0.011f4 + 0.782f5 + η1

z15 = 0.895f1 + 0.245f2 + 0.017f3 + 0.273f4 + 0.080f5 + η15

where z1 is the standardized REACT variable and z15 is the standardized FAT variable

Interpretation of this model is not necessarily easy! Below is what Johnson first says about interpreting the common factors:

Interpretation of the rotated factors requires researchers to possess knowledge, experience, discretion, and wisdom, while remaining objective, and while suppressing, as much as possible, their own biases and prejudices. Researchers must carefully consider the population being sampled when making interpretations and should always keep in mind that the underlying factors are measuring unique and independent characteristics of the population that was sampled.

Please remember that the loadings for the common factors represent correlations between the original variables and the common factors. The farther these loadings are away from zero, the more of an association exists between the common factors and the original variables. To help then see which of these loadings are “away from zero”, one could set the cutoff argument value to something other than 0.0. Johnson uses 0.40, but this value could easily be set to something lower. Below is the output from using 0.40.

> print(x = mod.fit5v, cutoff = 0.4)

Call:

factanal(x = set1[, -1], factors = 5, rotation = "varimax")

Uniquenesses:

 REACT HEIGHT WEIGHT SHLDR PELVIC CHEST THIGH PULSE

 0.370 0.109 0.028 0.313 0.485 0.081 0.055 0.621

 DIAST CHNUP BREATH RECVR SPEED ENDUR FAT

 0.870 0.465 0.587 0.005 0.522 0.826 0.058

Loadings:

 Factor1 Factor2 Factor3 Factor4 Factor5

REACT 0.782

HEIGHT 0.888

WEIGHT 0.614 0.615 0.424

SHLDR 0.747

PELVIC 0.585

CHEST 0.488 0.458 0.666

THIGH 0.957

PULSE 0.575

DIAST

CHNUP -0.690

BREATH 0.598

RECVR 0.948

SPEED -0.534

ENDUR

FAT 0.895

 Factor1 Factor2 Factor3 Factor4 Factor5

SS loadings 3.149 2.843 1.760 0.948 0.905

Proportion Var 0.210 0.190 0.117 0.063 0.060

Cumulative Var 0.210 0.399 0.517 0.580 0.640

Test of the hypothesis that 5 factors are sufficient.

The chi square statistic is 53.8 on 40 degrees of freedom.

The p-value is 0.0712

Please remember that the common factors are independent of each other. Thus, each common factor needs to have a distinct interpretation. Below are possible interpretations of the common factors where I draw heavily upon Johnson’s thoughts about the data. A researcher could likely judge better why particular combinations of body size measurements make sense.

Factor 1: This may be a measurement of body size. In particular, this could be a measure of obesity level. The negative correlation with CHNUP would make sense because the larger one’s obesity level, the smaller number of chin-ups that one could complete (generally speaking). Again, hopefully, the subject matter researcher could make more sense of this.

Factor 2: This again could be a measurement of body size. In this case, it may be more geared toward skeletal structure due to the inclusion of variables like height and pelvic.

Factor 3: Cardiovascular fitness level

Factor 4: This again could be a measurement of body size. Johnson acknowledged difficulty with interpreting it. He suggests In this case, it may be with respect to a “measure of upper body strength” and whether or not an individual lifts weights.

Factor 5: Reaction time

The fifth common factor represents only one original variable, REACT, and it does not appear in any of other common factors. It may make sense to consider this variable separately from the other variables.

Johnson also points out that DIAST and ENDUR do not appear as a substantial component of any common factor. Thus, he suggests that the data is truly 7-dimensional.

As a reminder, the judgments above are a based on using a cut-off of 0.4 for the common factor loadings. Other cut-offs could lead to different interpretations. This is a problem with a factor analysis and other explanatory analysis methods. A key then is to state your assumptions and limitations for your analysis.

* 1. Examine and interpret the appropriate plots of the factor scores.

Obviously, simple scatter plots will not work well here due to the number of common factors. Instead, we could use plots that allow for a larger number of dimensions.

Examining the factor scores could be used in situations where one wants to identify applicants with desirable characteristics. For example, maybe large common factor #1 values would be desirable. In that case, applicant #15 may be the most desirable:

> factor.scores[factor.scores[,2] == max(factor.scores[,2]),]

 ID Factor1 Factor2 Factor3 Factor4 Factor5

15 15 2.483271 0.2377963 1.902411 -0.3149539 0.07053817