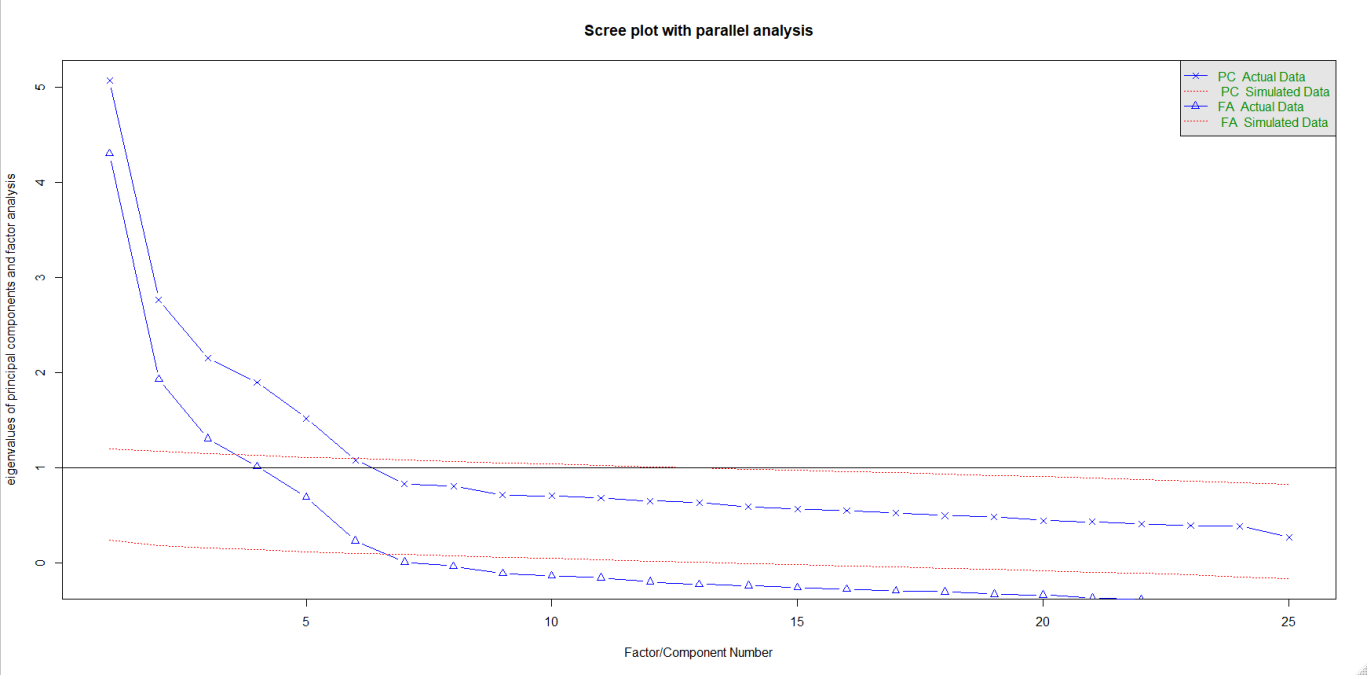
Assignment 2

May 2021

0)

There seems to be enough data to do an EDA with 80 subject items, 4 times the 20 number rule. I noticed that there seems to be a strong correlation between the variables in the group with N4 and A3 having strong correlations. It seems the beginning variables or early numbered variables 2,3,4 have high correlations and as the higher the number goes the correlation increases. We can view the correlation by listing out the variables in an x and y variable and compare them against one another to show how strong they are correlated to one another.

1)



From the Scree plot, we should retain 6 factors. To account for 90% overall variability we should retain the first 19 factors. Using the eigenvalue .+1 rule, we should retain the 4 factors that have eigenvalues over 1.

2)

ML1 = 0.51*A2 + 0.63*A3 + 0.63*A5 – 0.52*E1 – 0.57*E2 + 0.57*E3 – 0.73*E4

ML2 = 0.81*N1 + 0.79*N2 + 0.72*N3 +0.51*N4 + 0.51*N5

ML3 = 0.54*C1 + 0.63*C2 + 0.57*C3 – 0.67*C4 – 0.57*C5

ML4 = -0.53*O5

```
Call:
factanal(factors = 4, covmat = bfi_cor, n.obs = 2236, rotation = "varimax")

Uniquenesses:
   A1    A2    A3    A4    A5    C1    C2    C3    C4    C5    E1    E2    E3    E4    E5    N1    N2    N3    N4    N5    O1    O2    O3    O4    O5
0.946 0.721 0.610 0.742 0.575 0.673 0.607 0.681 0.509 0.577 0.721 0.582 0.531 0.462 0.627 0.346 0.373 0.471 0.591 0.697 0.678 0.715 0.511 0.867 0.713

Loadings:
   Factor1 Factor2 Factor3 Factor4
A1 -0.196   0.124
A2  0.509           0.141
A3  0.615           0.109
A4  0.422           0.218  -0.167
A5  0.631  -0.143
C1                  0.528   0.202
C2                  0.607   0.102
C3                  0.555
C4          0.225  -0.654
C5 -0.172   0.267  -0.567
E1 -0.517                  -0.104
E2 -0.590   0.219  -0.106  -0.102
E3  0.607                   0.308
E4  0.716  -0.127
E5  0.464           0.309   0.246
N1          0.805
N2          0.787
N3          0.723
N4 -0.269   0.549  -0.185
N5          0.516          -0.173
O1  0.198           0.108   0.520
O2          0.180  -0.118  -0.483
O3  0.319                   0.620
O4          0.191           0.301
O5                         -0.523

               Factor1 Factor2 Factor3 Factor4
SS loadings      3.263   2.670   1.989   1.553
Proportion Var   0.131   0.107   0.080   0.062
Cumulative Var   0.131   0.237   0.317   0.379

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 2631.66 on 206 degrees of freedom.
The p-value is 0
```

The cutoff value I used for deciding which loadings were sufficiently large for interpretation is +/- .5. The proportion of overall variability that is explained by the model 41.5% which is a good start but could be possibly improved upon if we increased the amount of variables from the 4 we are using. The statistical inference suggest we don't have the correct number of factors to describe the correlation matrix and thus we should increase it. We cannot reject the null hypotheses.

3)

This models has better interpretability than the task 2 model because using the .5 rule, we have the same number of variables than the model with the varimax. No the statistical inference for the maximum likelihood factor analysis does not suggest the model has the correct number of factors to describe the correlation matrix especially because the p-value is low suggesting that we add more factors.

4)

It is possible to find the correct number of factors but it would take a lot of experimentation with the different variables and rules for eigenvalues. The cutoff we use for deciding the loadings is the same as the previous models which is .5. The model that is easiest to interpret is k=1 that has 6 coefficients that pass the .5 cutoff however we could add more factors to this. The models that represent the correct numbers based on the inference results is k=10 which has a large p-value results and thus we still cannot reject the null hypotheses.

5)

Our easiest to interpret model compares well with the BIF model. The results from our model show that Neuroticism is expressed as the first factor, conscientiousness with the third factor, and openness with the fifth factor. Interestingly, we have a larger amount of factors than we have in previous models.

6)

There are gender differences and it seems to show different correlations based upon gender. It is more difficult to determine if personality is related to education. Here the model doesn't show much statical significance and the model accuracy is low. It is possible that with an improve accuracy score the model could show different results, but that is unclear. For the last part, the model shows the weakest correlation are very limited correlation to personality and age. Only two of the five factors show statistical significance and the model had a poor accuracy score in correctly predicting ages. Nevertheless, it seems like it is a weak correlation.

7)

This assignment was tasked with EDA on different BFI or personality related data. The data was very interesting as I had never worked with this before and the final question of finding the correlation based on gender, age, and education. It was interesting to test the factors against those three independent factors to see what were the results. In devising our EDA we took the data and constructed a correlation matrix. We then applied the eigenvalue rule to act as a cut off for the dataset. We were able to eliminate some of the factors and create various models based on the rule. The models were then further explored to see what information we could glean from them. Here, we came to understand the scree plot and the varying eigenvalues which explained variance. We fine tuned the model to only show statistically significant factors and reduced the factors to the number which provided the highest p-value. This was an interesting yet fulfilling assignment in which the correlation of data values was explored.