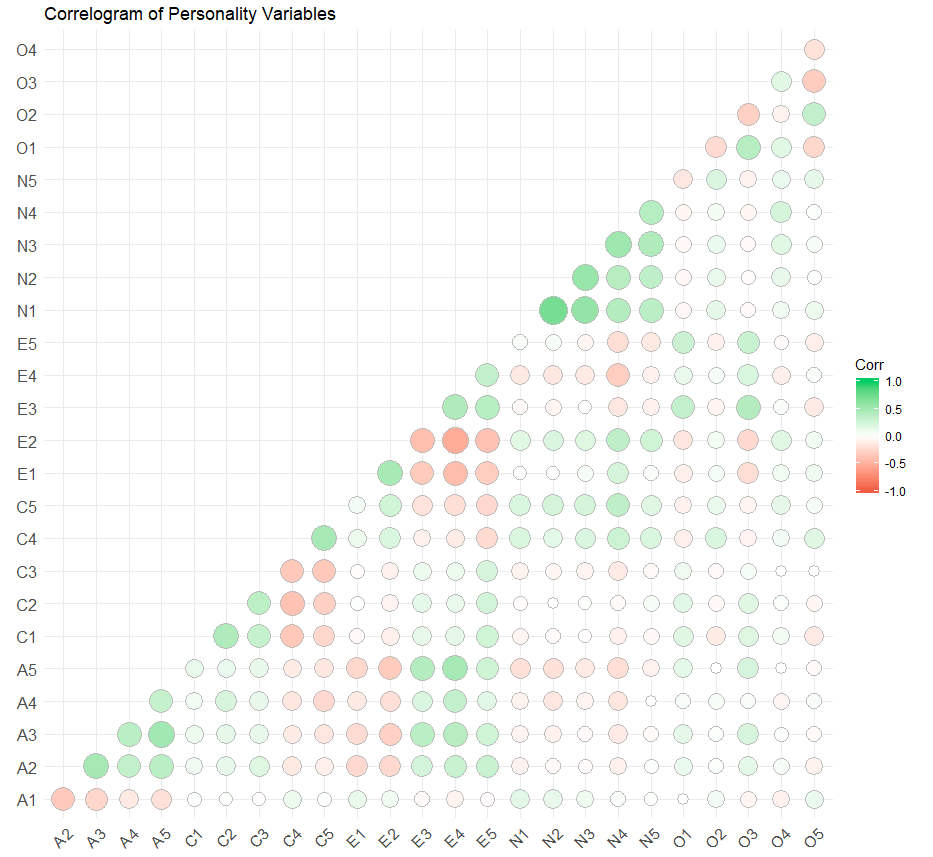
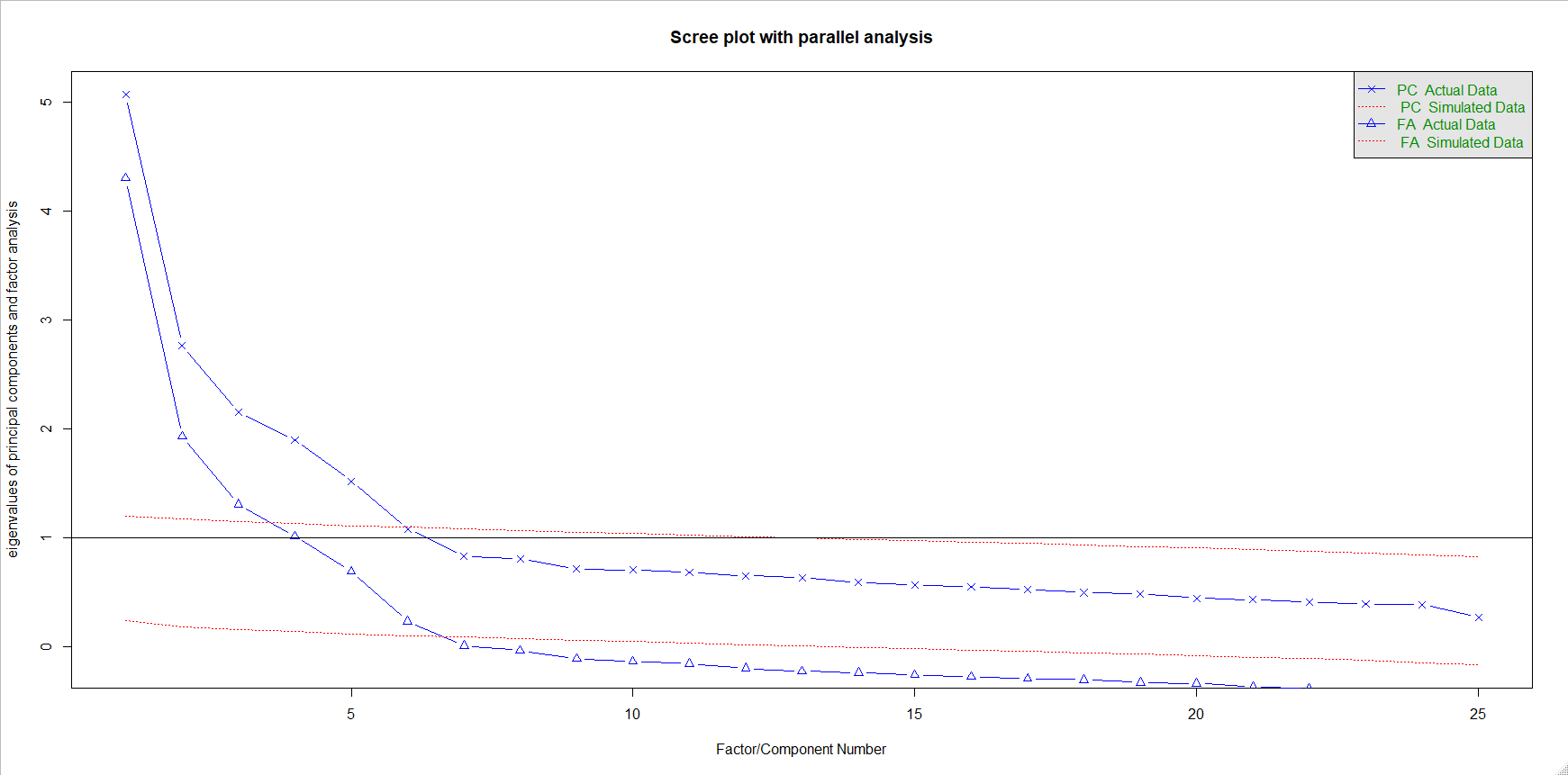
Assigment #2

Brandon Moretz

### Tasks

1. Conduct a basic Exploratory Data Analysis of this data.
   1. Is there enough data to conduct a basic Exploratory Factor Analysis on this data? Use the 20 times number of variables rule of thumb to decide.
      1. Yes, there is enough data here. There are approximately 80 subjects per item under study, which well exceeds the 20 rule.
   2. Obtain the correlation matrix for the 25 personality variables. What do you notice about the correlations? Are there any discernable patterns just looking at the correlation matrix?
      1. There seems to be a somewhat strong lagging correlation between the variables in the groups. We can see it pronounced along the main diagonal below:
      2. There are additional pockets of strong high/low correlations scattered throughout, although not quite as pronounced as the main diagonal.
2. Obtain the eigenvalues and eigenvectors of the correlation matrix.
   1. How many factors should you retain using the scree plot rule?
      1. From the Scree plot, we should retain 6 factors.
   2. How many factors should you retain to account for 90% of the overall variability?
      1. We would need to keep the first 19 factors to account for **90.7**% of the total variability in the data.
   3. How many factors should you retain using the eigenvalue >= 1 rule?
      1. There are 4 factors with eigenvalues over 1.
3. Use the eigenvalue >= 1 rule for the number of factors to retain.
   1. Estimate a factor model for the number of factors with eigenvalues greater than 1.
      1. This model is for the standard factor analysis without rotation, using only the 4 factors with eigenvalues >= 1.

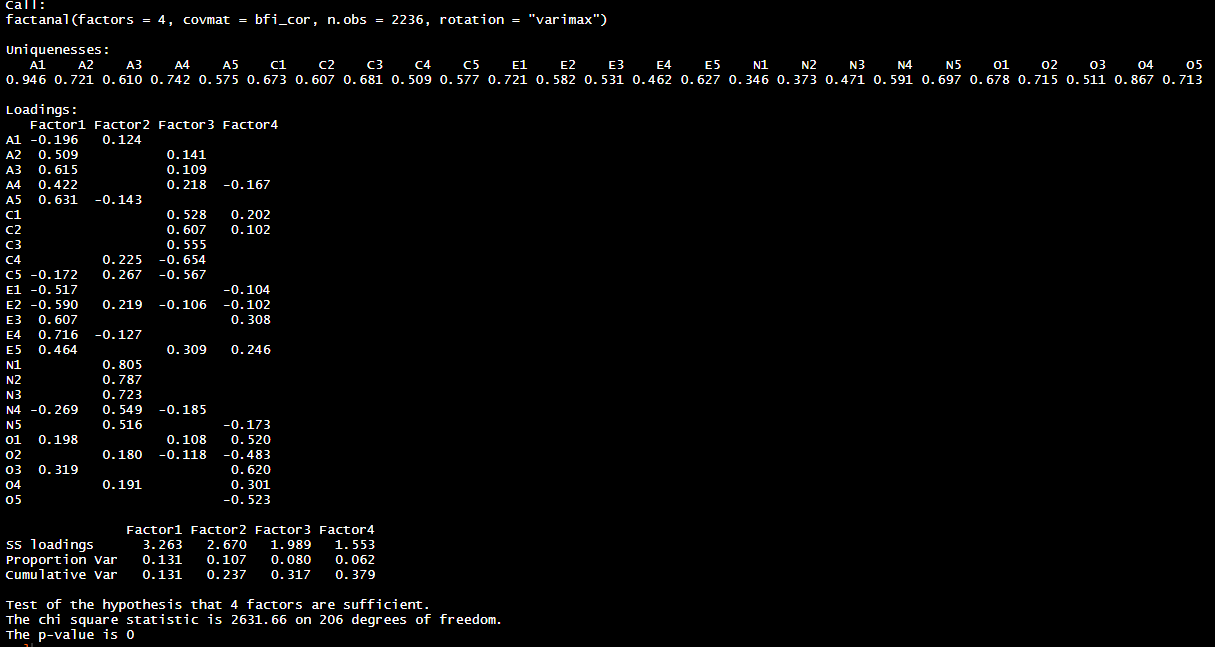
ML1 = 0.51\*A2 + 0.63\*A3 + 0.63\*A5 – 0.52\*E1 – 0.57\*E2 + 0.57\*E3 – 0.73\*E4

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

* 1. Use maximum likelihood factor analysis with a VARIMAX rotation. Report the factor loadings table and interpret each factor:



ML1 = 0.51\*A2 + 0.61\*A3 + 0.63\*A5 – 0.52\*E1 – 0.59\*E2 + 0.67\*E3 + 0.72\*E4

ML2 = 0.81\*N1 + 0.79\*N2 + 0.72\*N3 +0.55\*N4 + 0.52\*N5

ML3 = 0.51\*C1 + 0.52\*C2 + 0.57\*C3 – 0.67\*C4 – 0.57\*C5

ML4 = -0.52\*O5

* 1. What cutoff value did you use for deciding which loadings were sufficiently large for interpretation?
     1. +/- 0.5
  2. What proportion of overall variability is explained by this model? Is that sufficient to you?
     1. The total variability in the data explained by the factors is 41.5%, which is quite low even though we are only using 4 factors as opposed to the 25 underlying variables in our model to reduce the complexity. I think we should look to add more factors in follow-up analysis.

1. The VARIMAX factor rotation is an example of an orthogonal factor rotation. We also have oblique factor rotations. One example of an oblique factor rotation is the PROMAX rotation. Fit the same model from Task 2) but this time use the PROMAX rotation using maximum likelihood factor analysis.
   1. Does this model have better interpretability than the Task 2 Model with the VARIMAX rotation?
      1. It looks like in this case, using the same .5 threshold rule, we would have the same number of variables for interpretation that we had in the varimax rotation.
   2. Does the statistical inference for this maximum likelihood factor analysis suggest that this model has the correct number of factors to describe this correlation matrix? Should the factor rotation affect the statistical inference for the number of factors?
      1. Here, the statistical significance that 4 factors were sufficient to describe the data is very low. The statistical significance did change with the rotation, however, with all 3 models have low p-values in the hypothesis test that 4 factors are sufficient, indicating we are not using enough factors to describe the data.
2. Can we find the correct number of factors to describe this correlation matrix? Fit factor models using a VARIMAX rotation for k=1 through max(number of factors to retain from task 1 computations). For each factor model fit, use the factor loadings to interpret the individual factors.
   1. What cutoff value did you use for deciding which loadings were sufficiently large for interpretation?
      1. We will still use the .5 threshold for loading interpretation.
   2. Some of these will be easier to interpret than others. Which model is the easiest to interpret.
      1. The easiest model to interpret is k=1, which has a total of 6 coefficients that meet the .5 cutoff in the only factor. However, it is not statistically significant that 1 factor is sufficient to describe this data.
   3. Do any of these models represent the correct number of factors based on the inference results?
      1. If we use k=10 factors, we get a large enough p-value that we would fail to reject the null hypothesis that 10 factors are enough to describe the data. Although, we should note the factors after 5 have eigenvalues that are less than 1, and the final three factors have no coefficients that are above the .5 threshold, so we will exclude them from the final model.
   4. The final model with 11 factors (8 interpreted) would be:

F1 = 0.822\*N1 + 0.835\*N2 +0.708\*N3 + 0.515\*N4

F2 = 0.664\*A3 + 0.651\*A5 + 0.501\*E3 + 0.547\*E4

F3 = 0.519\*C1 + 0.504\*C2 + 0.558\*C3 -0.663\*C4 – 0.605\*C5

F4 = 0.644\*E1 + 0.674\*E2

F5 = -0.516\*O2 + 0.619\*O3 – 0.576\*O5

F6 = -0.511\*A1 + 0.604\*A2

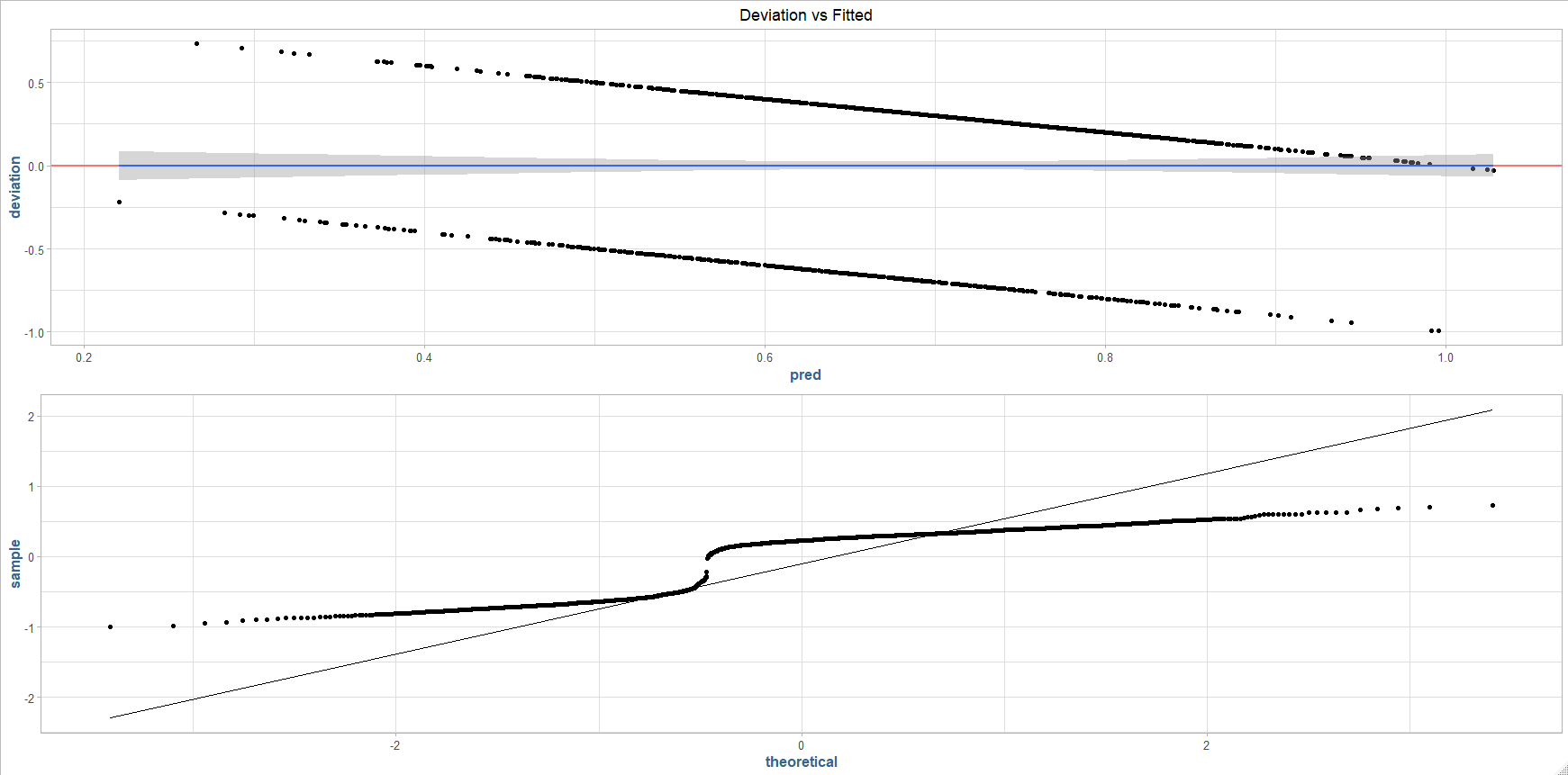
F7 = 0.823\*C2

F8 = 0.525\*N5

1. The researchers who commissioned the BFI data collection had a theory about personalities. According to their theory, there are 5 factors contained in this data. They are: Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Opennness. The variable naming convention (A, C, E, N, O) indicates which variables should band together to measure the associated latent trait. How does your easiest to interpret or best fitting model from Task 4) compare to this structure?

* 1. Yes, these groups make sense from what we have observed in the data. The correlation matrix from the beginning of this analysis indicated that the letter grouped variables tended to be highly correlated with each other. From the model derived in the previous exercise, we can clearly see that Neuroticism is expressed with the first factor, Conscientiousness with the third factor, Openness with the fifth factor. Others are more dispersed throughout the factors, however, we have over twice as many factors in the previous model as we do underlying latent traits / associations.

1. Just to be certain, refit a 5-factor model using the VARIMAX rotation and maximum likelihood factor analysis. Save the Factor Scores as variables to the BFI dataset. Use the Factors and response variables to determine:
   1. Are there gender differences?

Using the factor scores from the 5-factor model to construct a logistic regression model to predict gender, we can see there is a clear correlation of the factor scores to gender:

The model derived from the 5-factors for gender is:

Gender = 0.227 + 0.036F1 + 0.33F2 + 0.036 F3 + 0.052F4 -0.064F5

Where F1-5 are the factors derived from the previous analysis.

The model predicted the gender for males 96.45% of the time, as opposed to only 12.5% for females. Given the large discrepancy in these results, we would likely conclude that there are gender differences amongst the factors.

* 1. If personality is related to education
  2. If personality types are related to age

What do you conclude?

### Conclusion