ASSIGMENT #2
BRANDON MORETZ

## TASKS

1.) Conduct a basic Exploratory Data Analysis of this data.

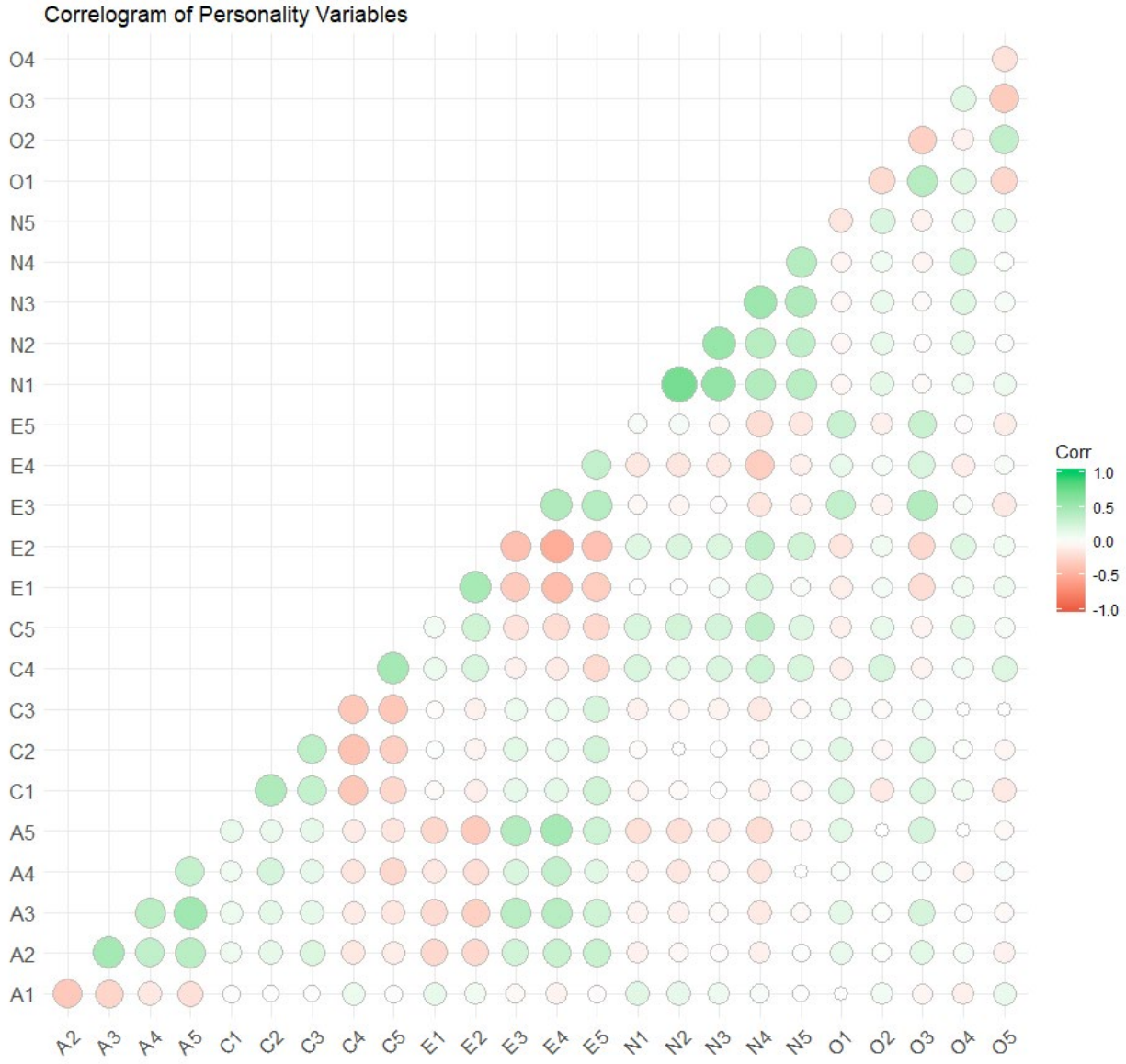
a. Is there enough data to conduct a basic Exploratory Factor Analysis on this data? Use the 20 times number of variables rule of thumb to decide.

i. Yes, there is enough data here. There are approximately 80 subjects per item under study, which well exceeds the 20 rule.

b. 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?

c.

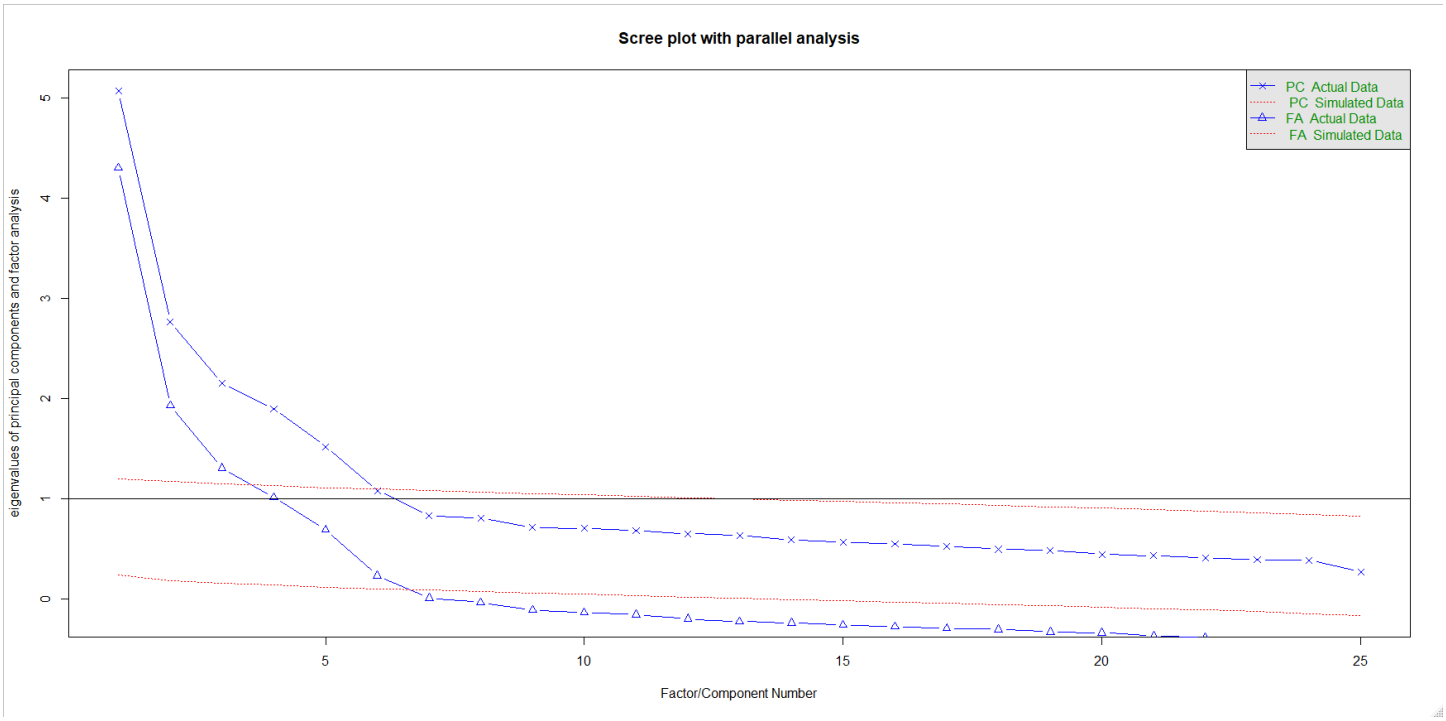
i. There seems to be a somewhat strong lagging correlation between the variables in the groups. We can see it pronounced along the main diagonal below:




ii. 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.

a. How many factors should you retain using the scree plot rule?




i. From the Scree plot, we should retain 6 factors.

b. How many factors should you retain to account for 90% of the overall variability?

i. We would need to keep the first 19 factors to account for **90.7**% of the total variability in the data.

c. How many factors should you retain using the eigenvalue >= 1 rule?

i. There are 4 factors with eigenvalues over 1.

3.) Use the eigenvalue >= 1 rule for the number of factors to retain.

a. Estimate a factor model for the number of factors with eigenvalues greater than 1.

i. This model is for the standard factor analysis without rotation, using only the 4 factors with eigenvalues >= 1.

$ML_1 = 0.51*A2 + 0.63*A3 + 0.63*A5 - 0.52*E1 - 0.57*E2 + 0.57*E3 - 0.73*E4$
$ML_2 = 0.81*N1 + 0.79*N2 + 0.72*N3 + 0.51*N4 + 0.51*N5$
$ML_3 = 0.54*C1 + 0.63*C2 + 0.57*C3 - 0.67*C4 - 0.57*C5$
$ML_4 = -0.53*O5$

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

```
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
```

$ML_1 = 0.51*A2 + 0.61*A3 + 0.63*A5 - 0.52*E1 - 0.59*E2 + 0.67*E3 + 0.72*E4$

$ML_2 = 0.81*N1 + 0.79*N2 + 0.72*N3 + 0.55*N4 + 0.52*N5$

$ML_3 = 0.51*C1 + 0.52*C2 + 0.57*C3 - 0.67*C4 - 0.57*C5$

$ML_4 = -0.52*O5$

c. What cutoff value did you use for deciding which loadings were sufficiently large for interpretation?
   i. +/- 0.5

d. What proportion of overall variability is explained by this model? Is that sufficient to you?

   i. 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.

4.) 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.

a. Does this model have better interpretability than the Task 2 Model with the VARIMAX rotation?
   i. 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.

b. 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?
   i. 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.

5.) 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.

   a. What cutoff value did you use for deciding which loadings were sufficiently large for interpretation?
      i. We will still use the .5 threshold for loading interpretation.

   b. Some of these will be easier to interpret than others. Which model is the easiest to interpret.
      i. 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.

   c. Do any of these models represent the correct number of factors based on the inference results?
      i. 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.

   d. The final model with 11 factors (8 interpreted) would be:

$F_1 = 0.822*N1 + 0.835*N2 + 0.708*N3 + 0.515*N4$
$F_2 = 0.664*A3 + 0.651*A5 + 0.501*E3 + 0.547*E4$
$F_3 = 0.519*C1 + 0.504*C2 + 0.558*C3 - 0.663*C4 - 0.605*C5$
$F_4 = 0.644*E1 + 0.674*E2$
$F_5 = -0.516*O2 + 0.619*O3 - 0.576*O5$
$F_6 = -0.511*A1 + 0.604*A2$
$F_7 = 0.823*C2$
$F_8 = 0.525*N5$

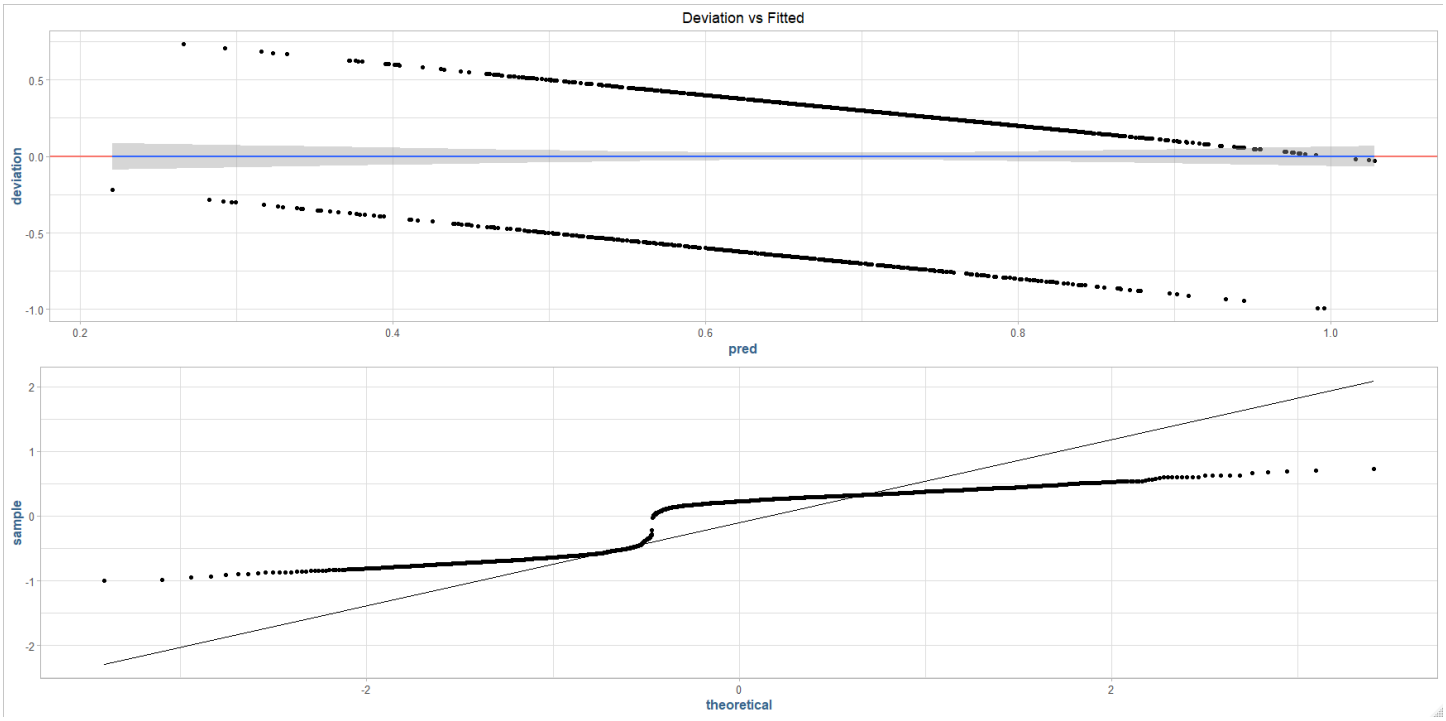
6.) The researchers who commissioned the BFI data collection had a theory about personalities. According to their theory, there are 5 factors contained in this data. They are: Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness. The variable naming convention (A, C, E, N, O) indicates 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?

   a. 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.

7.) 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:

a. 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:

$$\text{Gender} = 0.227 + 0.036F_1 + 0.33F_2 + 0.036\,F_3 + 0.052F_4 - 0.064F_5$$

Where $F_{1\text{-}5}$ are the factors derived from the previous analysis.

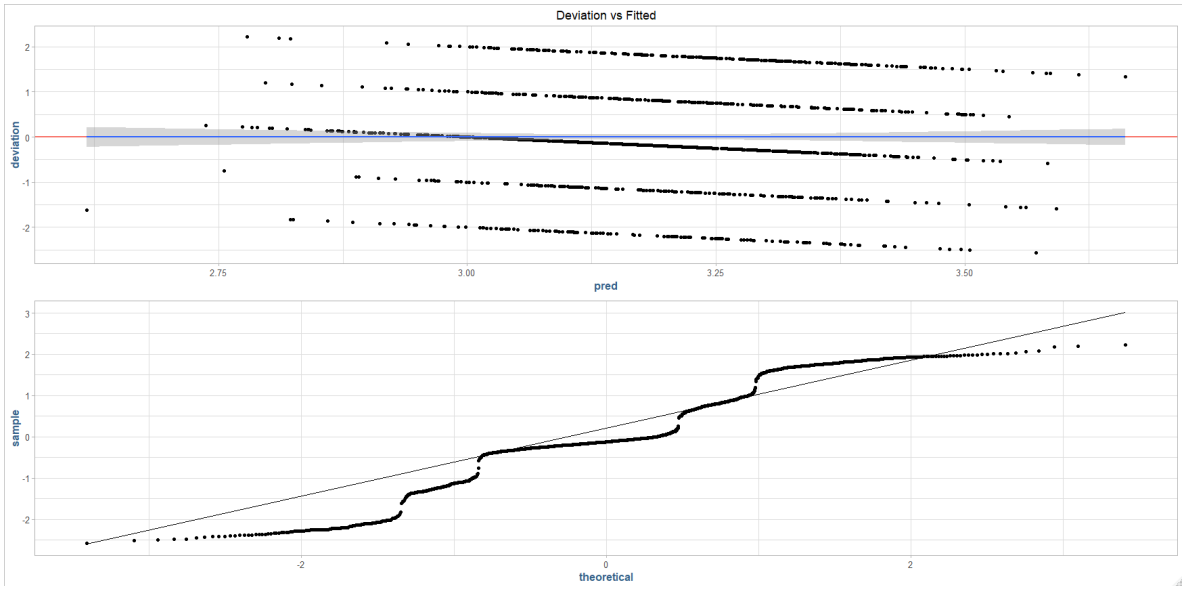
The model predicted the gender for males 96.45% of the time, as opposed to only 12.5% for females. Given the large discrepancy in these results, we would likely conclude that there are gender differences amongst the factors. All the coefficients in this model showed statistical significance, and the model produced an MSE of .2, which is relatively low.

b. If personality is related to education

Using a standard linear model to predict education from the five factor scores, we derive the following model:

$$\text{Education} = 3.027 - 0.15F_1 - 0.029F_2 + 0.018F_3 - 0.017F_4 + 0.129F_5$$

Where $F_{1\text{-}5}$ are the respective factor scores.




The preceding plot shows the residual diagnostics for the education model. We should note that none of the coefficients in this model shows statistical significance, and the models MSE is 1.2, which is relatively high for this scale of data. The model accuracy is only 41%, and the scores are highly concentrated in the ~3 range. From the results of this model, I wouldn't say there is a strong relationship between the factors scores and education.

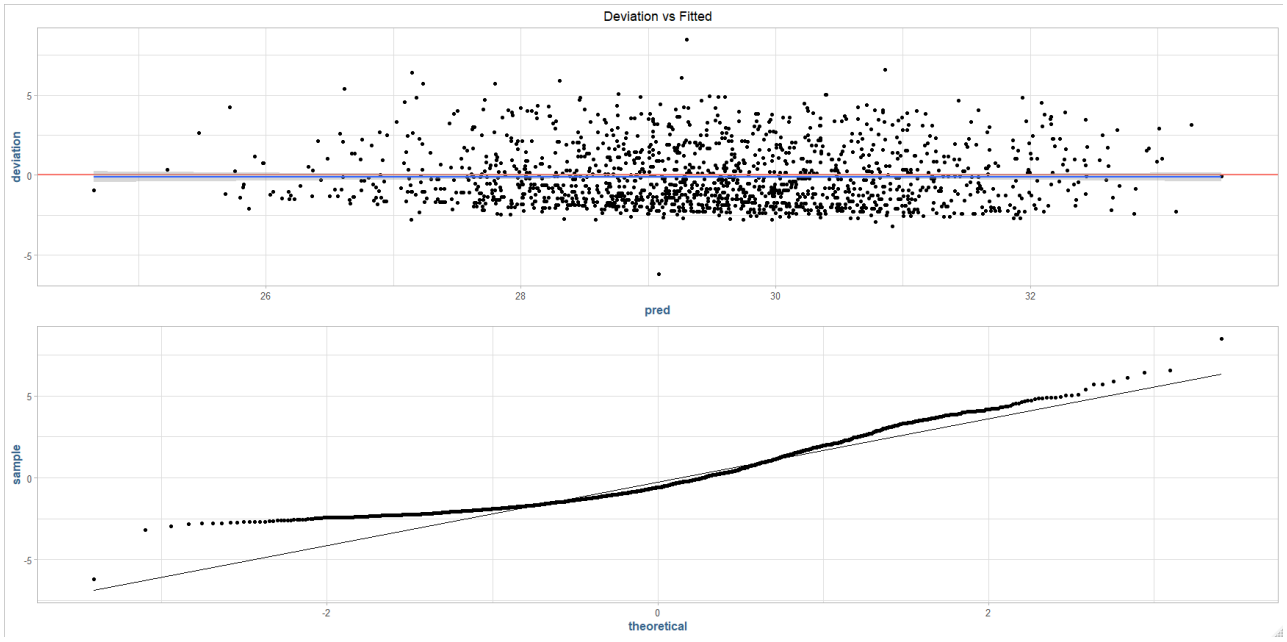
c. If personality types are related to age

For the age variable, we fit a generalized linear model with a quasi-Poisson distribution due to the age variables distribution in the underlying data. The model fit for the age variable is:

$$Age = 3.3 - 0.19F_1 + 0.007F_2 + 0.023F_3 - 0.016F_4 + 0.014F_5$$

Where $F_{1\text{-}5}$ are the respective factor scores.

The model regression diagnostics are:



We note that only two of the five factor coefficients show statistical significance, and overall the model only predicted approximately 2% of the ages correctly from the factor data.

This is the weakest modeled relationship thus far, and we would conclude there is no relationship between the age variable and the personality factors.

## CONCLUSION

This lab has been an interesting introduction to exploratory factor analysis. We took a dataset consisting of 25 personally attributes and looked for underlying relationships in the dataset. We did note that the variables that start with the same letter in seemed to be at least somewhat correlated amongst all the groups, given the large dots in the main diagonal of the correlogram. We constructed various factor models from these correlations which were derived from various "factor cut-off rules", all of which seem to be subjective in nature.

During the exploration of the various models, we did note that the "statistically significant" attribute of the model does appear to have some issues. For instance, most of the rules for picking the number of factors from the Scree plot to eigenvalues and explained variance produced wildly varying estimates, and then models derived from these estimates were deemed not to be "statistically significant". In the end when learned the true number of "latent traits" (which, was kind of shown to us due to the naming and the correlogram), we generated a model that was not "significant", however, did produce some interpretable results.

The process of drawing statistical inferences from these factors was a bit obtuse and difficult to interpret due to the large number of variables that are "mixed-in" to produce the final factors. Overall, a this was a incredibly informative lab on the overall process if exploratory factor analysis.