

|  |
| --- |
| Assignment #2: Exploratory Factor Analysis  *MSDS 411* |

Data: The data for this assignment comes from the International Personality Item Pool (ipip.ori.org) as part of the Synthetic Aperture Personality Assessment (SAPA) web based personality assessment project. The BFI data consists of the 25 personality self reported items (i.e. survey questions) obtained from 2800 subjects. Three additional demographic variables (sex, education, and age) are also included. This data is freely available in the PSYCH package of the R-Project system.

You can use the following code to obtain, load, and see the original data:

*install.packages("psych")*

*library(psych)*

*bfi\_data=bfi*

*bfi\_data*

The personality variables in the BFI data set are all Likert type variables measured on a scale from 1 to 6. Each variable is based on a statement, where the values for the variable are: 1 = not at all like me, and 6=totally like me. The statements and codes associated with each variable are:

|  |  |
| --- | --- |
| A1 Am indifferent to the feelings of others.  A2 Inquire about others' well-being.  A3 Know how to comfort others.  A4 Love children.  A5 Make people feel at ease.  C1 Am exacting in my work.  C2 Continue until everything is perfect.  C3 Do things according to a plan.  C4 Do things in a half-way manner.  C5 Waste my time.  E1 Don't talk a lot.  E2 Find it difficult to approach others.  E3 Know how to captivate people.  E4 Make friends easily.  E5 Take charge. | N1 Get angry easily.  N2 Get irritated easily.  N3 Have frequent mood swings.  N4 Often feel blue.  N5 Panic easily.  O1 Am full of ideas.  O2 Avoid difficult reading material.  O3 Carry the conversation to a higher level.  O4 Spend time reflecting on things.  O5 Will not probe deeply into a subject.  Demographic variables:  Gender (Males = 1, Females =2)  Education (1 = HS, 2 = finished HS, 3 = some college, 4 = college graduate 5 = graduate degree)  Age (age in years) |

## Source

The items are from the ipip (Goldberg, 1999). The data are from the SAPA project (Revelle, Wilt and Rosenthal, 2010), collected Spring, 2010 ( [https://sapa-project.org](https://sapa-project.org/)).

## References

Goldberg, L.R. (1999) A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In Mervielde, I. and Deary, I. and De Fruyt, F. and Ostendorf, F. (eds) Personality psychology in Europe. 7. Tilburg University Press. Tilburg, The Netherlands.

Revelle, W., Wilt, J., and Rosenthal, A. (2010) Individual Differences in Cognition: New Methods for examining the Personality-Cognition Link In Gruszka, A. and Matthews, G. and Szymura, B. (Eds.) Handbook of Individual Differences in Cognition: Attention, Memory and Executive Control, Springer.

Revelle, W, Condon, D.M., Wilt, J., French, J.A., Brown, A., and Elleman, L.G. (2016) Web and phone based data collection using planned missing designs. In Fielding, N.G., Lee, R.M. and Blank, G. (Eds). SAGE Handbook of Online Research Methods (2nd Ed), Sage Publcations.

# Assignment Tasks:

1. Conduct a basic Exploratory Data Analysis of this data. You will notice that there are missing values indicated by NA’s. To make things simple, only retain the data points that have complete information.

*#Remove rows with missing values and keep only complete cases*

*bfi\_data=bfi\_data[complete.cases(bfi\_data),]*

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.

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?

You will want to save the correlations as a matrix. If you can figure out how to do this directly, great. If you don’t know, or can’t find it quickly, then you’ll have to do it by hand. Here is some code to help. First, load the correlation matrix into R. How do we do that? Start with a vector of values, and then read that vector of values into a matrix object. Here is an example to help you figure out what to do for your data.

*cor.values <- c(1.000,0.210,0.370,-0.32,0.000,-0.31,-0.26,0.090,-0.38,*

*0.210,1.000,0.090,-0.29,0.120,-0.30,-0.14,0.010,-0.39,*

*0.370,0.090,1.000,-0.31,-0.04,-0.30,-0.11,0.120,-0.39,*

*-0.32,-0.29,-0.31,1.00,-0.16,0.25,-0.13,-0.14,0.900,*

*0.00,0.120,-0.04,-0.16,1.000,-0.20,-0.03,-0.08,-0.38,*

*-0.31,-0.30,-0.30,0.25,-0.20,1.000,-0.24,-0.16,0.180,*

*-0.26,-0.14,-0.11,-0.13,-0.03,-0.24,1.000,-0.20,0.040,*

*0.090,0.010,0.120,-0.14,-0.08,-0.16,-0.20,1.000,-0.24,*

*-0.38,-0.39,-0.39,0.900,-0.38,0.180,0.040,-0.24,1.000*

*);*

*# How do we put these correlation values into a correlation matrix?;*

*help(matrix)*

*cor.matrix <- matrix(cor.values,nrow=9,ncol=9,byrow=TRUE);*

*# Check that object is a matrix object;*

*is.matrix(cor.matrix)*

*# Check that matrix is symmetric;*

*# This check helps check for typos;*

*isSymmetric(cor.matrix)*

We can check most data types in R using an is.\* function. We type cast in R using an as.\* function.

1. Obtain the eigenvalues and eigenvectors of the correlation matrix. You can obtain this information in a number of different ways. You could use direct matrix functions or you could use the fa() function in the PSYCH package. Also, the Classroom may have other ways – check those resources. The goal for this task is to obtain a scree plot to go along with the eigenvalues. How many factors should you retain using the scree plot rule? How many factors should you retain to account for 90% of the overall variability? How many factors should you retain using the eigenvalue >= 1 rule?
2. Use the eigenvalue >= 1 rule for the number of factors to retain. Estimate a factor model for the number of factors with eigenvalues greater than 1. Use maximum likelihood factor analysis with a VARIMAX rotation. Report the factor loadings table and interpret each factor. What cutoff value did you use for deciding which loadings were sufficiently large for interpretation? What proportion of overall variability is explained by this model? Is that sufficient to you? You can use the fa() function of the PSYCH package or factanal() from the base STAT system.

factors\_data <- fa(r = cor\_matrix, nfactors = 6)

factors\_data <- factanal(covmat=cor\_matrix, n.obs=1442, factors=3, rotation='varimax');

names(f.1)

Does the statistical inference for the maximum likelihood factor analysis suggest that you have the correct number of factors to describe this correlation matrix? What is the null hypothesis for the chi-square test statistic? Do we reject or fail to reject this null hypothesis? Note that this hypothesis cannot be expressed in statistical notation like most hypotheses tests in Predict 410. (Hint: See Section 11.5 of Everitt.)

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?
   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?
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. What cutoff value did you use for deciding which loadings were sufficiently large for interpretation? Some of these will be easier to interpret than others. Which model is the easiest to interpret. Do any of these models represent the correct number of factors based on the inference results?
3. 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?
4. 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. If there are gender differences
   2. If personality is related to education
   3. If personality types are related to age

What do you conclude?

1. Please write a reflection on your experiences.

# Assignment Document:

All assignment reports should answer each of the questions separately. Please be sure to clearly indicate which question is being addressed. Results should be presented and discussed in an organized manner with the discussion in close proximity of the results. The report should not contain unnecessary R-code, intermediary computations, R-results, or non-essential information. The document should be submitted in pdf format. Name your file Assign2\_LastName.pdf.