# Quantitative Methods of Business Research Seminar 4

The goal of the fourth seminar is to study the procedures of exploratory factor analysis and clustering.

For the factor analysis, use dataset SAQ.sav. It is devoted to questionnaire for measuring ‘R anxiety’ and it is designed to measure various aspects of students’ anxiety towards learning R, the SAQ, and includes 23 questions. Each question was a statement followed by a 5-point Likert scale: ‘strongly disagree’, ‘disagree’, ‘neither agree nor disagree’, ‘agree’ and ‘strongly agree’ (SD, D, N, A, and SA, respectively). The questionnaire should help to know whether anxiety about R could be broken down into specific forms of anxiety. In other words, what latent variables contribute to anxiety about R? It includes 2571 completed questionnaires and this example is fictitious.

To apply cluster analysis, use BurgersOriginal.sav. This data set contains information on all kinds of characteristics of burgers from different restaurants.

**Protocol for the Seminar 4  
Factor analysis**

Q1. Preliminary analysis of correlations.

1. Use R-matrix to comment on weak correlations on your data (r < 0.3).
2. Comment on possible multicollinearity in the data (if r > 0.9 for any variables).
3. Check the determinant. Is it ok?

1. Check the results of Bartlett’s test of sphericity. What conclusions can you make?

1. Are there any variables that may cause problems? Should you exclude any variables before procced with factor analysis? What variables?

Q2. Sampling adequacy for multiple and individual variables.

1. What is the value for the overall Kaiser-Meyer-Olkin statistic? What does it mean?
2. Check that KMO values for individual variables are greater than 0.5. Made conclusions on the results.

Q3. Factor extraction.

1. Is Kaiser’s criterion applicable for this data? Why?
2. How many factors should be extracted according to Kaiser’s criterion (eigenvalues > 1)?
3. Check the scree plot? How many factors should be extracted according to it?
4. Conclude on your final decision about the number of factors to retain in the model.

Q4. Factor rotation.

1. Comment on orthogonal rotation (varimax). What factors could you spot? Try to give them the names.
2. Comment on oblique rotation (oblimin). Are there any differences between the results for different rotations?
3. Using Factor correlation matrix (for oblique rotation) comment on what rotation is more suitable for your data (it depends upon whether factors extracted are correlated).

Q5. Reliability analysis and final reports.

1. Report the results of your factor analysis.

**Example report:**

A principal axis factor analysis was conducted on the 23 items with oblique rotation (direct oblimin). The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis, KMO = 0.93, and all KMO values for individual items were greater than 0.77, which is well above the acceptable limit of 0.5. Four factors had eigenvalues over Kaiser’s criterion of 1 and in combination explained 50.3% of the variance. The scree plot was ambiguous and showed inflexions that would justify retaining either 2 or 4 factors. We retained 4 factors because of the large sample size and the convergence of the scree plot and Kaiser’s criterion on this value. Table 1 shows the factor loadings after oblique rotation. The items that cluster on the same factor suggest that factor 1 represents a fear of statistics, factor 2 represents peer evaluation concerns, factor 3 a fear of computers, and factor 4 a fear of math.

Table 1. Summary of exploratory factor analysis results for the R anxiety questionnaire (N =2571) after oblique rotation (Direct oblimin method)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item | Fear of statistics | Fear of PC | Fear of math | Fear of peep review |
| Statistics makes me cry | **0,54** | 0,19 | 0,24 | -0,01 |
| Standard deviations excite me | **-0,49** | -0,17 | -0,15 | 0,42 |
| I dream that Pearson is attacking me with correlation coefficients | **0,55** | 0,26 | 0,23 | -0,04 |
| I don't understand statistics | **0,47** | 0,24 | 0,17 | -0,05 |
| I weep openly at the mention of central tendency | **0,58** | 0,24 | 0,22 | -0,16 |
| I can't sleep for thoughts of eigenvectors | **0,43** | 0,04 | 0,09 | -0,24 |
| I wake up under my duvet thinking that I am trapped under a normal distribution | **0,57** | 0,27 | 0,15 | -0,18 |
| People try to tell you that R makes statistics easier to understand but it doesn't | **0,52** | 0,37 | 0,10 | -0,17 |
| I have little experience of computers | 0,06 | **0,77** | 0,13 | -0,10 |
| All computers hate me | 0,38 | **0,56** | 0,15 | -0,14 |
| Computers are useful only for playing games | 0,17 | **0,37** | 0,12 | -0,12 |
| I worry that I will cause irreparable damage because of my incompetence with computers | 0,31 | **0,55** | 0,23 | -0,15 |
| Computers have minds of their own and deliberately go wrong whenever I use them | 0,41 | **0,47** | 0,14 | -0,14 |
| Computers are out to get me | 0,30 | **0,36** | 0,24 | -0,20 |
| R always crashes when I try to use it | 0,39 | **0,60** | 0,13 | -0,14 |
| I slip into a coma whenever I see an equation | 0,33 | 0,26 | **0,63** | -0,05 |
| I have never been good at mathematics | 0,24 | 0,14 | **0,76** | 0,00 |
| I did badly at mathematics at school | 0,25 | 0,26 | **0,70** | -0,18 |
| My friends are better at statistics than me | -0,12 | -0,07 | 0,06 | **0,57** |
| My friends will think I'm stupid for not being able to cope with R | -0,18 | -0,03 | 0,00 | **0,49** |
| Everybody looks at me when I use R | -0,27 | -0,14 | -0,06 | **0,38** |
| My friends are better at R than I am | -0,04 | -0,15 | -0,07 | **0,44** |
| If I'm good at statistics my friends will think I'm a nerd | 0,03 | -0,04 | -0,07 | **0,31** |
| Eigenvalues | 7,3 | 1,7 | 1,3 | 1,2 |
| % of variance | 11,7 | 8,1 | 7,5 | 5,8 |
| Cronbach’s α | 0.82 | 0.82 | 0.82 | 0.57 |

1. Comment on Cronbach’s alpha values for all scales. For each subscale check Item-total statistics (use Corrected Item-Total Correlation): check for items that don’t correlate with the overall score from the scale. Are there any items in your subscales that probably should be dropped?
2. Report on reliability analysis.

**Example report**

The ‘Fear of computers’, ‘Fear of statistics’ and ‘Fear of maths’ subscales all had high reliabilities, all Cronbach’s α = 0.82. However, the ‘Fear of peer review’ subscale had relatively low reliability, Cronbach’s α = 0.57. Probably the ‘Fear of peer review’ subscale needs to be reconstructed.

**Cluster analysis**

Q6. The task is to cluster burgers using Calories, Sodium, Total Fat, and Sugars. At first standardize the values of the variables.

1. Insert the resulting Descriptive statistics table.
2. Which variable has the largest standard deviation before transformation?

Q7. Apply K-means with 3 clusters and 10 iterations. Use Sandwich values as labels. How many clusters should be built according to 30 indices?

1. Produce a bar for final cluster centers.
2. Report on final number of cases in each cluster.
3. Check that ***sugar\_g*** variable has significant differences for different clusters. Apply One-Way ANOVA for mean comparison of multiple groups.

Q8. Run Hierarchical Clustering to the same variables. Use Ward’s method and Squared Euclidean distance.

1. Conclude which clustering procedure is more convenient and understandable for you. Why?
2. Check that ***protein\_g*** variable has significant differences for different clusters. Apply One-Way ANOVA for mean comparison of multiple groups.