Multivariate Statistical Analysis Report #2

2016160037 이혜원

1. Introduction of factor analysis

Factor analysis is a statistical method to describe variability between observed variables and unobserved variables. Observed variables are called 'manifest variables' and unobserved variables that we want to measure is 'latent variables'. Factor analysis has similar format as regression analysis, but the difference is that in factor analysis model, manifest variables are regressed on unobserved variables. This analysis method is similar to PCA in a point that they use covariance matrix and they perform dimension reduction. There are several methods to do factor analysis; principal component method, principal factor method, maximum likelihood factor analysis.

2. Explanation of data

This data consists of student information data (ID, Gender), their guardians and 25 questions about their guardians and there are 52 observations in this data. There are three types of guardians, mother, father and others.

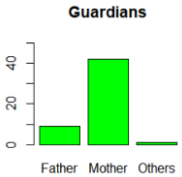
25 questions are either positive or negative questions about corresponding guardians, including vague questions. Questions 1, 3, 5, 6, 7, 11, 12, 15, 17, 21, 22, 25 seem to be questions related to positive aspects of guardians. On the other hand, questions 2, 4, 8, 9, 10, 13, 14, 16, 18, 19, 20, 23, 24 are negative questions about guardians. There are 4 answers to questions. The answer 'very like' has point 3, 'moderately like' has point 2, 'moderately unlike' has point 1, 'very unlike' has point 0. For negative questions, points are given in a reversed way. For example, 'very like' gets point 0, 'moderately like' gets 1, and so force.

3. Understand and explore the data

Before we explore the variables, we need to check if there are missing values in this data. There were two missing values in the data, in variable 'Q17' and 'Q23'. For better factor analysis, we need to delete the row that included missing data.

(1) Guardians

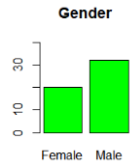
Now let's take a look at 'Guardians' variable first. This variables has 3 values, 'Father', 'Mother', 'Other guardian'. The 'Mother' has 42 observations, and 'Father' has 9. Below is the table and graph of distribution of variables.



Father	Mother	Other Guardian
9	42	1

(2) Gender

The 'Gender' variable has two values, 'Male' and 'Female'. Let's check its distribution. There are 12 male students more than female students.



Female	Male
20	32

(3) 25 Questions

The next step is to check how the answers to 25 questions are distributed. I divided questions into 2 groups as I mentioned in section 2 – positive and negative.

1) Positive questions: Q1, Q3, Q5, Q6, Q7, Q11, Q12, Q15, Q17, Q21, Q22, Q25

Most of the answers are positive, 'very like' or 'moderately like' rather than negative.

2) Negative questions: Q2, Q4, Q9, Q10, Q14, Q16, Q18, Q19, Q20, Q23, Q24

Most of the answers are negative, 'very unlike' or 'moderately unlike' rather than positive. This means that guardians were mostly positive to students.

4. Interpretation of factor analysis results

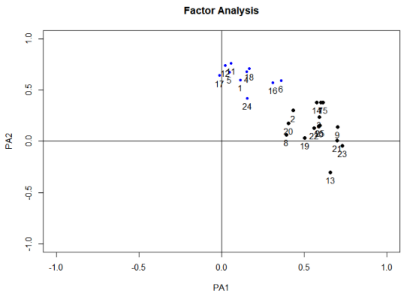
I applied two factor analysis methods; principal factor method and maximum likelihood factor analysis. Before conducting factor analysis, I replaced answers with numeric values and created correlation matrix.

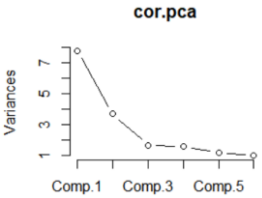
(1) Principal factor method

First, we have to choose how many factors we need to conduct analysis. According to scree plot of PCA of correlation matrix, 2 or 3 factors are needed. The third component is the elbow point of scree plot. Therefore, we are going to choose 2 factors in principal factor analysis. Apply 'varimax' rotation method to conduct analysis. The two factor loadings and plot of loadings are shown below.

Loadings

Variables	PA1	PA2		PA1	PA2
Q1	0.11	0.6	Q14	0.58	0.38
Q2	0.43	0.3	Q15	0.61	0.38
Q3	0.59	0.24	Q16	0.31	0.57
Q4	0.15	0.68	Q17	-0.01	0.64
Q5	0.05	0.67	Q18	0.17	0.71
Q6	0.36	0.59	Q19	0.5	0.03
Q7	0.6	0.38	Q20	0.41	0.18
Q8	0.39	0.06	Q21	0.7	0.01
Q9	0.7	0.14	Q22	0.56	0.13
Q10	0.59	0.15	Q23	0.73	-0.04
Q11	0.06	0.76	Q24	0.15	0.42
Q12	0.02	0.74	Q25	0.6	0.15
Q13	0.66	-0.3			



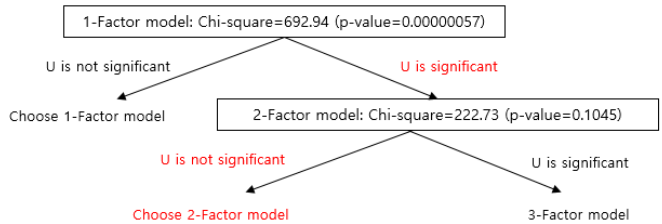


We can group 25 questions into 2 groups by comparing two loadings. For example, for Q1, loading of PA2 is bigger than PA1, so Q1 belong to PA2. This result of grouping is shown in graph above. PA1 explains guardian's effect on students' independency and PA2 explains their effect on student's feelings or emotions.

Attitude toward student's independency(PA1)	Attitude toward student's emotion(PA2)
2, 3, 7, 8, 9, 10, 13, 14, 15, 19, 20, 21, 22, 23, 25	1, 4, 5, 6, 11, 12, 16, 17, 18, 24

(2) Maximum likelihood factor analysis

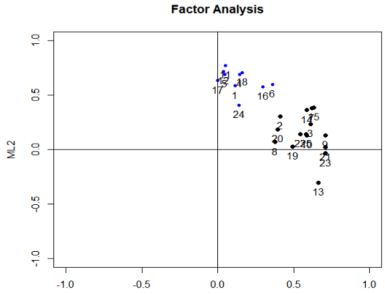
We start from number of factors from 1, and decide whether to increase factors or not in each step by checking p-value and chi square statistics. It's appropriate to use 2 factors for this data. This process is summarized below.



Now, conduct 2-factor maximum likelihood factor analysis, then we can get two factor loadings of 25 questions.

Loadings

Variables	ML1	ML2		ML1	ML2
Q1	0.11	0.58	Q14	0.59	0.36
Q2	0.41	0.3	Q15	0.63	0.38
Q3	0.61	0.23	Q16	0.3	0.68
Q4	0.14	0.69	Q17	0	0.63
Q5	0.05	0.69	Q18	0.16	0.71
Q6	0.36	0.6	Q19	0.49	0.03
Q7	0.62	0.38	Q20	0.39	0.18
Q8	0.38	0.07	Q21	0.71	0.02
Q9	0.71	0.13	Q22	0.54	0.14
Q10	0.58	0.13	Q23	0.71	-0.04
Q11	0.05	0.77	Q24	0.14	0.41
Q12	0.04	0.72	Q25	0.58	0.14
Q13	0.66	-0.3			



Similarly with principal factor analysis, 25 questions can be classified into 2 groups by comparing two loadings. ML1 explains guardian's effect on students' independency and ML2 explains their effect on student's feelings or emotions.

Attitude toward student's independency (ML1)	Attitude toward student's emotion(ML2)
2, 3, 6, 7, 8, 9, 10, 13, 14, 15, 19, 20, 21, 22, 23, 25	1, 4, 5, 11, 12, 16, 17, 18, 24

Compared to principal factor method, Q6 belongs to 'independency' group in maximum likelihood model. Q6 is 'was affectionate to me'. This question is vague in that whether the guardian is affectionate can be expressed in both behavior of teaching independency and emotion.

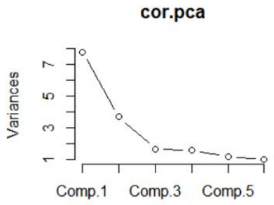
In both (1) and (2) factor analysis model, loadings are not similar, which gives clear interpretation of variables.

5. Factor analysis vs Principal component analysis

Conduct principal component analysis, then we can get following loadings of PCA. As we chose number of factors of factor analysis model by using scree plot, apply this to PCA then we are going to interpret 2 principal components. Let's take a look at the loadings we got in PCA.

PCA

Variables	Comp.1	Comp.2		Comp.1	Comp.2
Q1	0.18	0.219	Q14	0.253	
Q2	0.199		Q15	0.26	
Q3	0.224	-0.118	Q16	0.227	0.128
Q4	0.208	0.23	Q17	0.154	0.283
Q5	0.178	0.268	Q18	0.218	0.233
Q6	0.244	0.116	Q19	0.153	-0.183
Q7	0.258		Q20	0.162	
Q8	0.13	-0.131	Q21	0.196	-0.253
Q9	0.228	-0.196	Q22	0.191	-0.16
Q10	0.204	-0.157	Q23	0.191	-0.281
Q11	0.198	0.291	Q24	0.151	0.131
Q12	0.185	0.299	Q25	0.206	0.14
Q13	0.108	-0.364			



The first principal component explains overall scores each 25 question get, since components are similar. The second principal component contrasts Q1, 4, 5, 6, 11, 12, 16, 17, 18, 24, 25 and Q3, 8, 9, 10, 13, 19, 21, 22 and 23 by signs of loadings. Other questions are considered as not interpretable variables, vague questions. The first group is related to student's pleasure (emotions) and the second group is relate to student's independency. Therefore, we can conclude that interpretations of results of PCA and factor analysis are similar.

6. Factor analysis results depending on student's gender and different types of guardians

(1) Gender

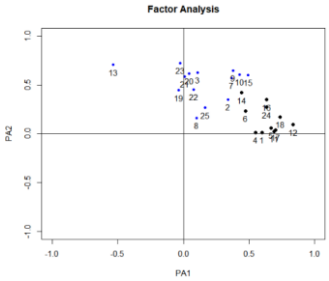
First, we have to divide data into 'Female' and 'Male' set. There are 20 Female, and 30 Male data in this data.

[Female]

According to scree plot of correlation matrix of Female, 2 factors seem to be appropriate for principal factor analysis. Then we can get those loadings of factors by conducting analysis.

Loadings

Variables	PA1	PA2		PA1	PA2
Q1	0.6	0.01	Q14	0.44	0.42
Q2	0.34	0.35	Q15	0.49	0.6
Q3	0.11	0.63	Q16	0.16	0.35
Q4	0.55	0.01	Q17	0.7	0.04
Q5	0.67	0.06	Q18	0.74	0.17
Q6	0.47	0.23	Q19	-0.04	0.44
Q7	0.37	0.57	Q20	0.04	0.62
Q8	0.1	0.16	Q21	0.01	0.58
Q9	0.38	0.65	Q22	0.08	0.45
Q10	0.43	0.61	Q23	-0.02	0.72
Q11	0.69	0.02	Q24	0.63	0.27
Q12	0.83	0.09	Q25	0.17	0.27
Q13	-0.54	0.71			



Questions can be classified into two groups, similar to previous factor analysis model.

Attitude toward student's independency (PA1)	Attitude toward student's emotion(PA2)
1, 4, 5, 6, 11, 12, 16, 17, 18, 24	3, 7, 9, 19, 20, 21, 22, 23

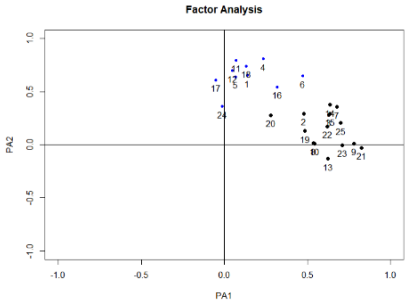
However, we cannot decide which group Q2, 8, 10, 13, 14, 15, 25 belongs to since their loadings do not have big differences.

[Male]

To conduct principal factor analysis, use 2 factors and rotate axis using 'varimax' method. Then loadings of factors are like below.

Loadings

Variables	PA1	PA2		PA1	PA2
Q1	0.14	0.65	Q14	0.64	0.37
Q2	0.48	0.29	Q15	0.63	0.29
Q3	0.63	0.28	Q16	0.32	0.55
Q4	0.24	0.81	Q17	-0.05	0.61
Q5	0.07	0.64	Q18	0.13	0.74
Q6	0.47	0.65	Q19	0.49	0.13
Q7	0.68	0.35	Q20	0.28	0.28
Q8	0.54	0.01	Q21	0.82	-0.03
Q9	0.78	0.01	Q22	0.62	0.17
Q10	0.54	0.01	Q23	0.71	-0.01
Q11	0.07	0.8	Q24	-0.01	0.36
Q12	0.05	0.7	Q25	0.7	0.2
Q13	0.62	-0.13			



Questions can be classified into two groups, similar to previous factor analysis model.

Attitude toward student's emotions (PA1)	Attitude toward student's independency (PA2)
2, 3, 7, 8, 9, 10, 13, 14, 15, 19, 21, 22, 23, 25	1, 4, 5, 11, 12, 16, 17, 18, 24

However, we cannot decide which group Q6, 20 belongs to since their loadings do not have big differences.

Male students' data was well-classified compared to that of Female students, which have many vague questions. Male students tend to answer 'very like/unlike' than female students. Moreover, Female students had vague questions such as Q2, 8, 10, 13, 14, 15, 25. Those questions are mostly about 'independency'. It seems that some of the guardians put more restrictions on female students' independency than they do on male students.

(2) Guardian

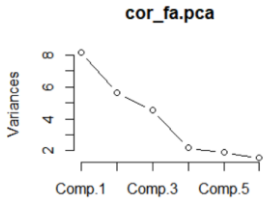
First, we have to divide data into 'Father' and 'Mother' sets. There are 8 observations in 'Father', and 41 observations in 'Mother'. Since there is only one data in 'Other guardians' option, we are not going to analyze it.

[Father]

There were only 8 observations in 'Father' dataset. However, there was an error when I tried to conduct factor analysis. So I tried PCA analysis for this data. According to scree diagram, 3 principal components seem to be useful for PCA.

PCA

Variables	Comp.1	Comp.2	Comp.3		Comp.1	Comp.2	Comp.3
Q1	0.245	0.155	0.247	Q14	0.264	0.116	-0.106
Q2		0.105	0.419	Q15	0.27		-0.182
Q3	0.149	0.346		Q16	0.127	-0.194	0.315
Q4	0.319			Q17	0.342		
Q5	0.279	-0.11		Q18		-0.249	0.23
Q6	0.319			Q19		0.338	0.197
Q7	0.259	0.238	-0.132	Q20	-0.178		0.111
Q8		-0.285	0.183	Q21		0.197	-0.188
Q9	0.264	0.238	-0.132	Q22		0.244	0.12
Q10		0.325	0.237	Q23	-0.21	0.142	
Q11	0.183	-0.206		Q24	0.104		0.395
Q12	0.279			Q25		0.338	0.232
Q13			-0.313				



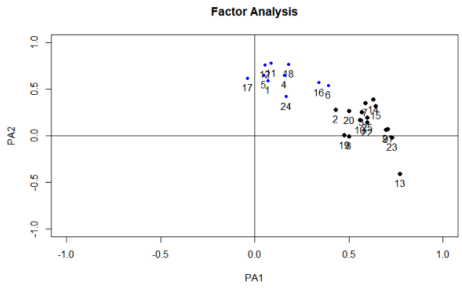
The first principal component explains overall score of the questions except for Q20 and Q23. The second principal component contrasts independency and emotions as other analysis does. But there are several vague questions (Q4, 6, 12, 13, 15, 17, 20, 24). We cannot find meaningful results by analyzing component 3.

[Mother]

We are going to use 2 factors to conduct analysis. Number of factors is decided by plotting scree diagram.

Loadings

Variables	PA1	PA2		PA1	PA2
Q1	0.07	0.59	Q14	0.63	0.39
Q2	0.43	0.28	Q15	0.64	0.32
Q3	0.57	0.25	Q16	0.34	0.57
Q4	0.16	0.65	Q17	-0.04	0.61
Q5	0.05	0.65	Q18	0.18	0.77
Q6	0.39	0.54	Q19	0.48	0
Q7	0.59	0.35	Q20	0.5	0.27
Q8	0.5	-0.01	Q21	0.71	0.07
Q9	0.7	0.06	Q22	0.6	0.14
Q10	0.56	0.17	Q23	0.73	-0.02
Q11	0.09	0.77	Q24	0.17	0.42
Q12	0.05	0.76	Q25	0.6	0.19
Q13	0.77	-0.41			



Questions can be classified into two groups. The result is similar to that of 'Male' of 'Gender' class.

Attitude toward student's independency (PA1)	Attitude toward student's emotion(PA2)
3, 7, 8, 9, 10, 13, 14, 15, 19, 20, 21, 22, 23, 25	1, 4, 5, 11, 12, 16, 17, 18, 24

Q2 and Q6 are vague questions (loadings of two factors are similar).

There were more vague questions in Father's cases. But it's hard to compare 'Father' and 'Mother' case since observations of 'Father' data were too small.

7. Conclusion

Factor analysis on data gives meaningful interpretation about the data. It's possible to classify variables into groups which have similar properties. Also we can easily check which one is the vague variable, especially Q6, which is hard to classify by comparing loadings of factors. Two methods, principal factor analysis and maximum likelihood analysis extracts similar number of factors and similar corresponding results. I personally think that ML method is more apparent since it decides number of factors to use based on p-value and chi-square statistics.

(※) Is factor analysis appropriate for given survey data?

Based on the results of two factor analysis and PCA, I concluded that factor analysis is useful for my data. Factor analysis classifies each question into groups that represents guardians' attitude toward students. I could easily catch what each group means in this data. Also, by comparing two factor analysis method, I could get 'vague questions' that is hard to decide whether it belongs to independency group or emotion group. Also as in section 6, we can compare characteristics of different types of Gender and Guardian data by apply factor analysis. Therefore, we can conclude that factor analysis is useful in comparing data and analyze the differences.