STAT401 Multivariate Statistical Analysis
Midterm Report (Factor Analysis)

2015100059, 통계학과, 윤창원

Ⅰ. Capturing the unobserved

In many studies, researchers try to indirectly capture variables that can not be observed through variables easily observed. Typical example of this would be relationship between test scores from certain subjects and student's various dimension of intelligence. This is where the factor analysis comes in. Factor analysis defines the observed variables as 'Manifest variables' and the unobserved variables as 'Latent variables' or just 'Factors' and sets relationship of form of multiple regression where Manifest variables are dependent variables and Latent variables are independent variables. That is, $x_i = \lambda_{i1}f_1 + \lambda_{i2}f_2 + \cdots + \lambda_{ik}f_k + u_i$ where, x_i is i-th manifest variable (i=1,...,p) and λ_{ij} is a coefficient(loading) for f_j which is j-th factor(j=1,...,k). Also, there is u_i which stands for residual for i-th manifest variable. Now, we can express this equation using matrices. $x = \Lambda f + u$. Here, x is a vector of x_i, Λ matrix contains loadings and μ is a vector of u_i. And from here, with some assumptions, we can easily derive that $\Sigma = \Lambda\Lambda' + \Psi$, where Σ is a covariance matrix of manifest variables and Ψ is diag(u_i). This equation decomposes covariance matrix into 'Communality' part ($\Lambda\Lambda'$) and 'Specific Variance' part (Ψ). From this relationship, we can estimate values of Λ through various methods such as Principal Component Method, Principal Factor Method and Maximum Likelihood Factor Method. The important fact is that the solution for Λ is not unique. So, we often find Λ in a manner that is easier to interpret the relationship which is called 'Simple Structure'. After estimating loadings, we try to match each variable to certain factor (with largest loading) and interpret the results.

Ⅱ. Data Exploration

Before diving into real analysis example, it is always a good start to explore the data we will use. The data is a survey data consisting of 25 Questions from 52 different individuals. Those questions are about the relationship between survey takers and their guardians in first 16 years. The answers should be chosen from 'Very likely', 'Moderately likely', 'Moderately unlikely' and 'Very unlikely'. Also, the data have personal info about the survey taker such as types of guardian('Mother', 'Father', 'Other') and gender along with ID.

First, the response to the questions were coded as follow to conduct further analysis. {'Very likely': 3, 'Moderately likely': 2, 'Moderately unlikely': 1, 'Very unlikely': 0} for positive questions and {'Very likely': 0, 'Moderately likely': 1, 'Moderately unlikely': 2, 'Very unlikely': 3} for negative questions marked (*).

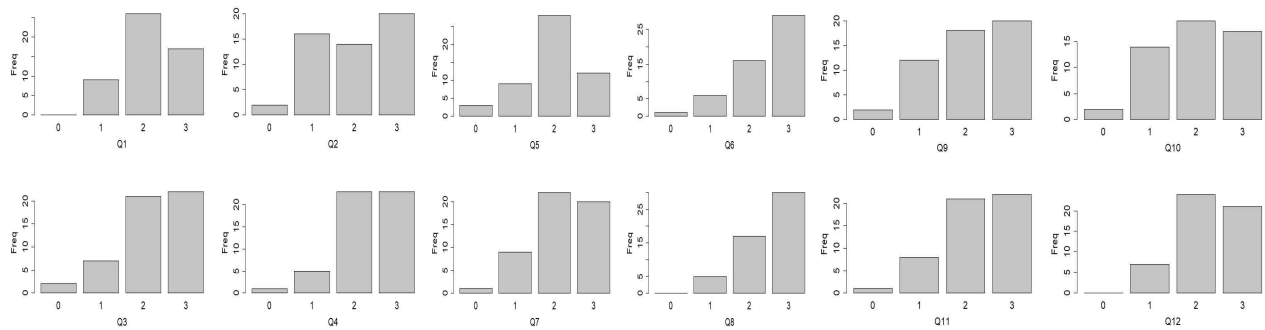
Numbers	Question
Q1	Spoke to me in a warm and friendly voice.
Q2 (*)	Did not help me as much as I needed.
Q3	Let me do those things I liked doing.
Q4 (*)	Seemed emotionally cold to me.
Q5	Appeared to understand my problems and worries.
Q6	Was affectionate to me.
Q7	Liked me to make my own decisions.
Q8 (*)	Did not want me to grow up.
Q9 (*)	Tried to control everything I did.
Q10 (*)	Invaded my privacy.
Q11	Enjoyed talking things over with me.
Q12	Frequently smiled at me.
Q13 (*)	Tended to baby me.
Q14 (*)	Did not seem to understand what I needed or wanted.
Q15	Let me decide things for myself.
Q16 (*)	Made me feel I wasn't wanted.
Q17	Could make me feel better when I was upset.
Q18 (*)	Did not talk with me very much.
Q19 (*)	Tried to make me feel dependent on her/him.
Q20 (*)	Felt I could not look after myself unless she/he was around.
Q21	Gave me as much freedom as I wanted.
Q22	Let me go out as often as I wanted.
Q23 (*)	Was overprotective of me.
Q24 (*)	Did not praise me.
Q25	Let me dress in any way I pleased.

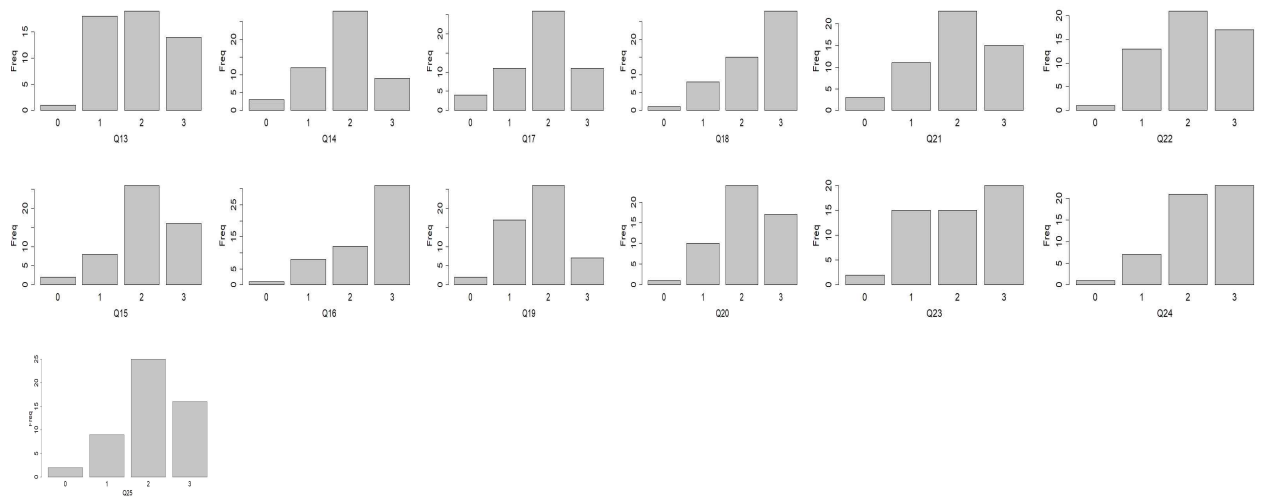
<Table 1>

The reason for this inverse coding for positive and negative questions is to set the same direction for each answers.

There was one missing value in Q17 and Q23 each. To impute for this missing values, the report used correlation matrix without the observations with missing value. Through the correlation matrix, found out that Q17 has high correlation with Q5, and Q23 has high correlation with Q9. So, for the missing value of Q17, assumed the survey taker would have answered the same with Q5 and used same method for the other missing value.

After this imputation, to take a deeper look of the data, plotted histograms of the answers for the questions.





<Figure 1>

Through histograms, one can easily identified that the data is left-skewed. Actually, when calculated the actual skewness of the data, every questions showed negative value. Hence, we can assume that the data is far from having normal distribution.

Ⅲ. Diving into real analysis

As in the PCA, one can choose whether to use covariance matrix or correlation matrix for the analysis. Since this is a survey data and each question is equally important, the report used correlation matrix to estimate the loadings.

The first method that will be used for estimation is Principal Factor Method.[1] Principal Factor Method uses similar approach to PCA and estimate Λ through spectral decomposition of $\Sigma^* = \Sigma - \Psi$. Here, choosing k is equivalent to choosing how many Principal Components to use.

For our data, when we conduct PCA, Proportion of Cumulative Variance and Kaiser's rule recommended 7 factors while Scree Plot recommended to use 3 factors. Since using 7 factors are too much for 25 variables, the report used 3 factors.

The other method is Maximum Likelihood method. This method needs strict assumption that the data's distribution is normal and estimate Λ through MLE. Hence, the result can be unstable for our data. Due to distributional assumption, we can test how many factors would be appropriate. For our data, it was 3.

Surprisingly, the result of grouping variables to specific factor was exactly the same for both method. For interpretation, Factor 1 captures Emotional Bonding between the survey taker and the guardian while Factor 2 captures Freedom that the survey taker could have. Lastly, Factor 3 captures dependency toward the guardian. However, the boundary between factor 2 and factor 3 is somehow very vague and this might be a signal that 2 factor model can be more interpretable.

---

1) Since Principal Component Method can be seen as a special case of Principal Factor Method, it will not be handled

<Principal Factor Method>

Factor 1	Q1, Q2, Q4, Q5, Q6, Q10, Q11, Q12, Q14, Q16, Q17, Q18, Q24
Factor 2	Q3, Q7, Q9, Q15, Q21
Factor 3	Q8, Q13, Q19, Q20, Q22, Q23, Q25

<Maximum Likelihood Method>

Factor 1	Q1, Q2, Q4, Q5, Q6, Q10, Q11, Q12, Q14, Q16, Q17, Q18, Q24,
Factor 2	Q3, Q7, Q9, Q15, Q21,
Factor 3	Q8, Q13, Q19, Q20, Q22, Q23, Q25

<Table 2 & 3>

Sometimes, Factor Analysis and Principal Components analysis are compared since both are seen as dimensionality reduction technics. Also for Principal Factor Method from FA, they share common methodology. Here are results from PCA for our data.

	PC1	PC2	PC3
Q1	0.203	0.118	near 0
Q2	0.168	near 0	near 0
Q3	0.214	-0.124	-0.298
Q4	0.240	0.134	0.285
Q5	0.230	0.198	-0.164
Q6	0.247	0.174	-0.244
Q7	0.217	-0.183	-0.338
Q8	0.144	-0.153	0.372
Q9	0.227	-0.203	-0.171
Q10	0.216	near 0	0.151
Q11	0.228	0.194	0.163
Q12	0.218	0.221	-0.151
Q13	0.109	-0.251	near 0
Q14	0.234	near 0	near 0
Q15	0.248	-0.165	-0.181
Q16	0.176	0.187	near 0
Q17	0.229	0.265	near 0
Q18	0.214	0.208	0.121
Q19	near 0	-0.287	0.163
Q20	0.145	-0.198	0.271
Q21	0.235	-0.264	-0.169
Q22	0.210	-0.222	0.227
Q23	0.167	-0.384	0.175
Q24	0.143	0.230	0.341
Q25	0.142	-0.114	near 0

<Table 4>

When we look at the PC2, it can be easily identified that PC2 exactly contrasts the Questions that were grouped with Factor 1 (Emotional Bonding) and questions grouped with Factor 1 and 2 (Freedom and Dependency). This result is a strong evidence for similarity between FA and PCA.

The other step of analysis will be conducting FA by groups divided by Gender and type of Guardian. Would they show any difference? For this analysis the report used Principal Factor Method and the number of factors were all recommended to be 3.

When we divide the group by gender, it showed a huge difference. The female group showed very similar result to the one with whole observation. However, the male group showed totally different result. Instead of having vague boundary between freedom and dependency, it linked related questions to Factor 1, dividing the once emotional bonding related questions in more detail. Here Factor 2 can be interpreted as emotional bonding as before and Factor 3 can be interpreted as an assist that the survey taker got from the guardian throughout first 16 years. But also, this makes the boundary between Factor 2 and Factor 3 vague.

<Male>

Factor 1	Q2, Q3, Q7, Q8, Q9, Q10, Q13, Q15, Q19, Q20, Q21, Q22, Q23, Q25
Factor 2	Q4, Q11, Q12, Q14, Q16, Q18,
Factor 3	Q1, Q5, Q6, Q17, Q24,

<Female>

Factor 1	Q1, Q2, Q4, Q5, Q6, Q10, Q11, Q12, Q14, Q16, Q17, Q18, Q24, Q25
Factor 2	Q8, Q13, Q15, Q19, Q20, Q22, Q23
Factor 3	Q3, Q7, Q9, Q21

<Table 5 & 6>

(* Q19 in Male had larger value of loading for Factor 3 in absolute value due to unstability from small obs, but selected among only positive values)

For the type of guardian, 42 survey takers out of 52 answered 'Mother'. Response of 'Father' only had 9 cases and 'Other Guardian' only had 1 case. Hence it was not possible to conduct FA for 'Father' group and 'Other Guardian' group due to lack of observations. So the report did not made comparison between type of guardians.

The final step is to conduct FA with two factors since 3 factor model continuously makes vague boundary between Factors.

Factor 1	Q1, Q2, Q4, Q5, Q6, Q10, Q11, Q12, Q14, Q16, Q17, Q18, Q24
Factor 2	Q3, Q7, Q8, Q9, Q13, Q15, Q19, Q20, Q21, Q22, Q23, Q25

<Table 7>

We can now see that the two factors can be clearly interpreted as 'Emotional Attachment' and 'Dependency toward the guardian'. Also it exactly matches with the result from PCA's PC2.

Ⅳ. Conclusion

Factor analysis is an excellent tool in that it helps us capture variables that we can not measure directly. It not only enables us to widen our research area but also is an useful dimensionality reduction method since it explains the data with much fewer common factors. In our example of data, we only needed 2 or 3 factors to explain our survey data and also nice and clear interpretation was possible. In comparison with PCA, Factor Analysis shares many similarity and as seen in the analysis, results of both method also shares some common parts. However, in the sense that Principal Components are mere recombination of the data we have, Factor Analysis would be more appropriate for our data since survey itself is usually conducted for a reason to capture the unobserved variables.

Even though the Factor Analysis is a powerful statistical method, it should be conducted carefully. It is better to check the distributional property of the data and also the number of factors to use should be thought thoroughly since the results can be quite different.