Multivariate Statistical Analysis Report 1

수학과 2016160037 이혜원

1. Introduction of Principal Component Analysis

   Principal component analysis (PCA) is a frequently used method in explanatory data analysis. The objective of PCA is to reduce dimensionality while retaining much of information without losing it. This process is done by creating uncorrelated variables from correlated variables, and those uncorrelated variables are called 'principal components'.

   Principal components are uncorrelated linear combinations of the variables in multivariate data. They are derived in decreasing order of importance. The first principal component is a component which accounts for most of the variation in data, and this means that it has the biggest variance. The second component has the biggest variance in remaining variation, and it's uncorrelated with the first principal component. We repeat this process several times until we get adequate amount of uncorrelated principal components.

   We decide the number of components to choose by using 3 methods. The first method is to check proportion of variance explained by components. Normally about 70-90% is suggested. The second method is to choose components whose eigenvalues are larger than average. If components are extracted from correlation matrix(R), we choose components whose eigenvalues are larger than 1. This is Kaiser's Rule. The last method is to use scree diagram. Scree diagram is the plot of eigenvalue against its index. We find 'elbow point', point where small eigenvalues begin. Then we choose the number of components equal or less than index of elbow point.

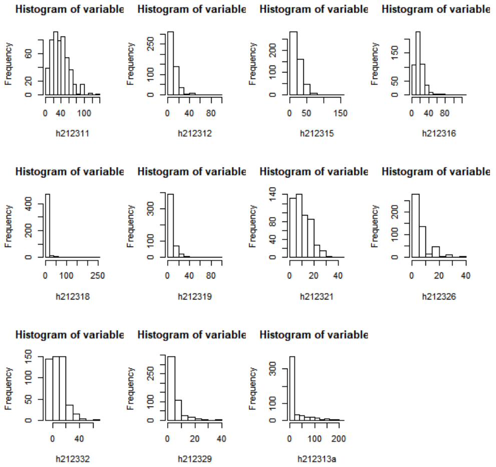
2. Explanation of data

   This data is about Korean labor market and income activities of households. Let's take a look at columns of this data.

   We are going to use 11 columns to conduct PCA analysis and those 11 columns are 'h212311, h212312, h212313a, h212315, h212316, h212318, h212319, h212321, h212326, h212332, h212329'. Those columns shows different types of living expenses such as food & groceries, meals out, education cost and so on. The rest of the columns give additional information about household. Id column is household ID. The 'h212102' column earned income and 'h212402' shows monthly savings amount. The 'age' and 'gender' columns show age and gender of household each.

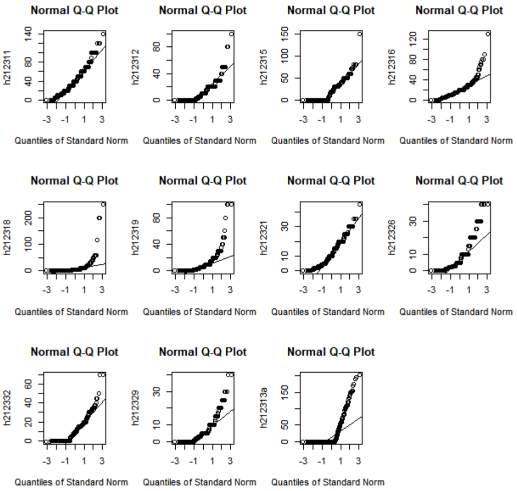
3. Exploring data with graphs and checking assumptions

   (1) Histogram of variables



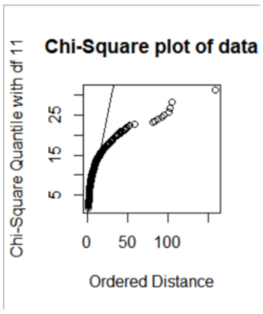
I drew histogram of each variables to check overall distribution.

(2) QQ-plot of variables



Only h212311 and h212321 variables seem to follow univariate normal distribution. Other variables deviates from the straight line so they do not follow normal distribution.

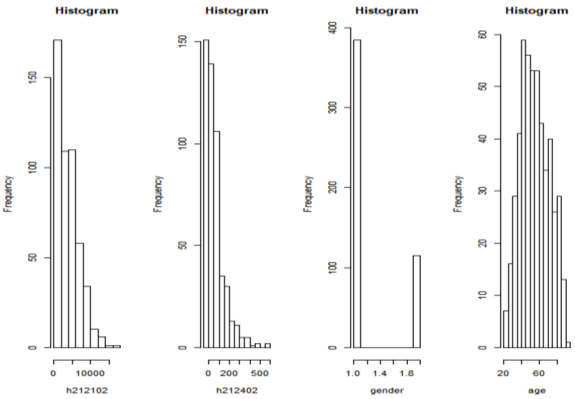
(3) Chi-square probability plot



This data does not follow multivariate normal distribution since it deviates from the straight line.

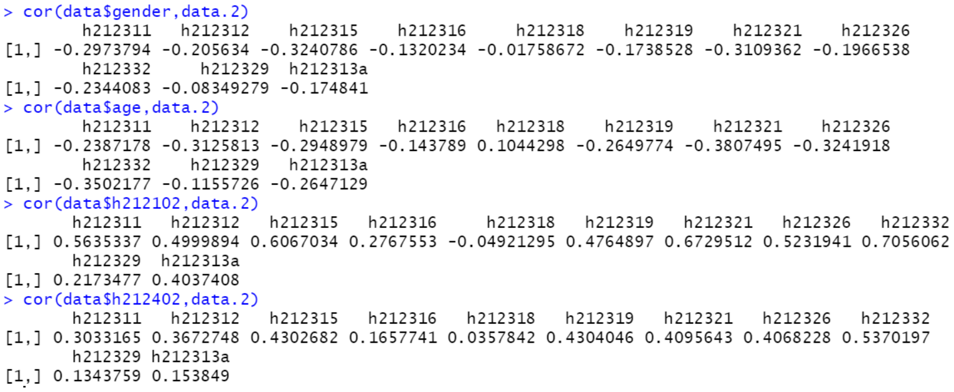
(4) Explore other variables

1) Histogram of remaining variables



‘Age’ variable seems to follow normal distribution but ‘h212102’ (earned income) and ‘h212402’ (monthly savings amount) does not. In ‘gender’ variable, proportion on ‘1’ is much bigger than ‘2’.

2) Question: are those variables have correlation with 11 living expense variables?



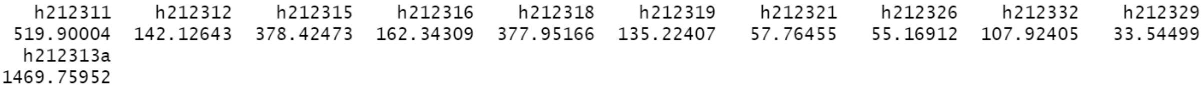
```
> cor(data$gender,data.2)
       h212311    h212312    h212315    h212316     h212318    h212319    h212321    h212326
[1,] -0.2973794 -0.205634 -0.3240786 -0.1320234 -0.01758672 -0.1738528 -0.3109362 -0.1966538
       h212332     h212329   h212313a
[1,] -0.2344083 -0.08349279 -0.174841
> cor(data$age,data.2)
       h212311     h212312    h212315    h212316   h212318     h212319    h212321    h212326
[1,] -0.2387178 -0.3125813 -0.2948979 -0.143789 0.1044298 -0.2649774 -0.3807495 -0.3241918
       h212332     h212329   h212313a
[1,] -0.3502177 -0.1155726 -0.2647129
> cor(data$h212102,data.2)
      h212311   h212312   h212315   h212316     h212318   h212319   h212321   h212326   h212332
[1,] 0.5635337 0.4999894 0.6067034 0.2767553 -0.04921295 0.4764897 0.6729512 0.5231941 0.7056062
      h212329  h212313a
[1,] 0.2173477 0.4037408
> cor(data$h212402,data.2)
      h212311   h212312   h212315   h212316   h212318   h212319   h212321   h212326   h212332
[1,] 0.3033165 0.3672748 0.4302682 0.1657741 0.0357842 0.4304046 0.4095643 0.4068228 0.5370197
      h212329 h212313a
[1,] 0.1343759 0.153849
```

Those are correlation between each remaining variable and 11 living expense variables. If you take a look at ‘h212102’ (earned income) case, it’s pretty highly correlated with communication costs and health insurance payment. But other variables’ (gender, age, h212402) correlation coefficients are mostly less than 0.5, so there is no remarkable correlation with 11 variables.

4. Principal component analysis about 11 living expense variables

I will use correlation matrix(R) for principal component analysis, since there are large differences among the variances of original variables so variables with large variances may dominate other components. If we use correlation matrix, all variables are equally important. Below are the results of PCA.

[Variance of each variable]

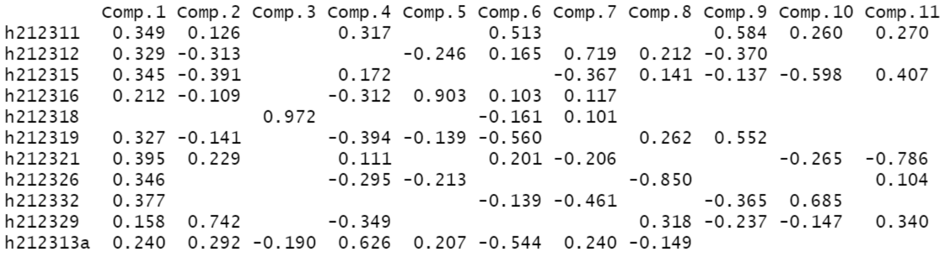


```
  h212311    h212312    h212315    h212316    h212318    h212319    h212321    h212326    h212332    h212329
519.90004  142.12643  378.42473  162.34309  377.95166  135.22407   57.76455   55.16912  107.92405   33.54499
 h212313a
1469.75952
```

5. Analysis PCA results

(1) Principal components

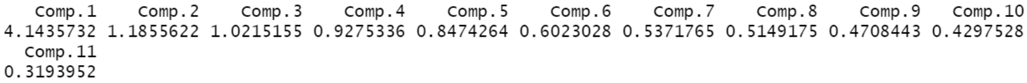
These are coefficients of each components.



```
         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11
h212311   0.349  0.126         0.317         0.513                0.584  0.260   0.270
h212312   0.329 -0.313               -0.246  0.165  0.719  0.212 -0.370
h212315   0.345 -0.391         0.172               -0.367  0.141 -0.137 -0.598   0.407
h212316   0.212 -0.109        -0.312  0.903  0.103  0.117
h212318                 0.972               -0.161  0.101
h212319   0.327 -0.141        -0.394 -0.139 -0.560         0.262  0.552
h212321   0.395  0.229         0.111         0.201 -0.206               -0.265  -0.786
h212326   0.346               -0.295 -0.213               -0.850                 0.104
h212332   0.377                             -0.139 -0.461        -0.365  0.685
h212329   0.158  0.742        -0.349                       0.318 -0.237 -0.147   0.340
h212313a  0.240  0.292 -0.190  0.626  0.207 -0.544  0.240 -0.149
```

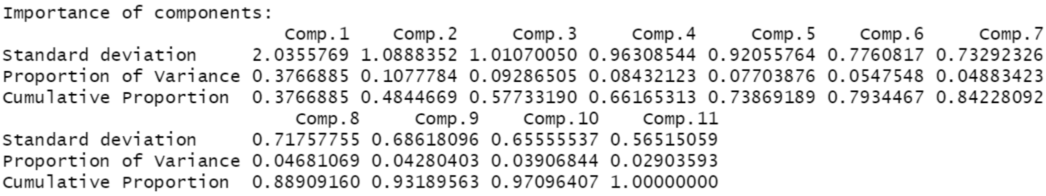
(2) Lambdas (variances)

By Kaiser's Rule, lambda of first three components is larger than 1, so 3 components seem important.



```
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    Comp.8    Comp.9   Comp.10 
4.1435732 1.1855622 1.0215155 0.9275336 0.8474264 0.6023028 0.5371765 0.5149175 0.4708443 0.4297528 
  Comp.11 
0.3193952 
```

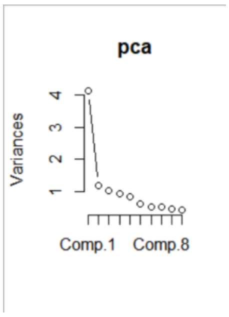
(3) Proportion of variance



```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5    Comp.6     Comp.7
Standard deviation     2.0355769 1.0888352 1.01070050 0.96308544 0.92055764 0.7760817 0.73292326
Proportion of Variance 0.3766885 0.1077784 0.09286505 0.08432123 0.07703876 0.0547548 0.04883423
Cumulative Proportion  0.3766885 0.4844669 0.57733190 0.66165313 0.73869189 0.7934467 0.84228092
                           Comp.8     Comp.9    Comp.10    Comp.11
Standard deviation     0.71757755 0.68618096 0.65555537 0.56515059
Proportion of Variance 0.04681069 0.04280403 0.03906844 0.02903593
Cumulative Proportion  0.88909160 0.93189563 0.97096407 1.00000000
```

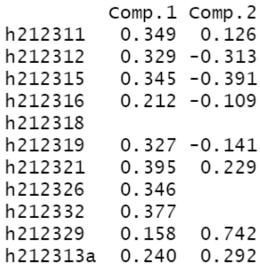
(4) Scree diagram

According to scree diagram, the elbow point is in 2nd component.



6. Interpretation of PCA result

By using 2 principal components, we can explain the data about 48.45%, and as I mentioned in the previous section, 2nd components is an elbow point. Therefore, we will interpret first two principal components.

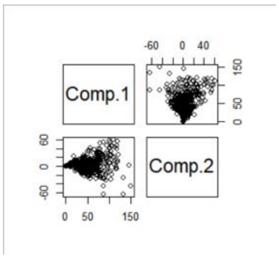


```
         Comp.1 Comp.2
h212311   0.349  0.126
h212312   0.329 -0.313
h212315   0.345 -0.391
h212316   0.212 -0.109
h212318              
h212319   0.327 -0.141
h212321   0.395  0.229
h212326   0.346       
h212332   0.377       
h212329   0.158  0.742
h212313a  0.240  0.292
```

Except for 'h212318' (health and medication cost), first principal component can explain overall living expenses since coefficients of 1st component are similar. 1st principal component has the biggest variance, which is 4.15 and it accounts for 37.67% of total variance.

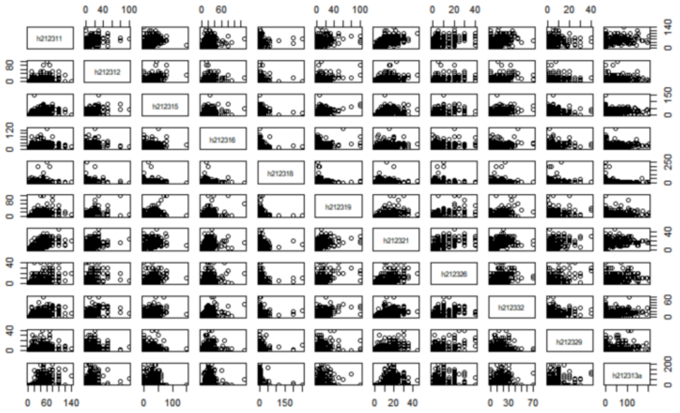
The second principal component contrasts ‘h212311, h212321, h212329, h212313a’ (food and groceries, communication costs, public transportation) and ‘h212312, h212315, h212316, h212319’ (meals out, vehicle maintenance, housing maintenance, recreation). Also, 2nd component explains ‘h212329’ (public transportation) most. But coefficients of ‘h212318, h212326, h212332’ (health and medical costs, purchase of clothing, health insurance payment) are small, so they are not included in interpretation. The 2nd principal component has variance 1.19 and it accounts for 10.78% of total variance.

This is the scatterplot matrix of two principal components. We can conclude that two principal components are uncorrelated, since this plot does not have patterns.



7. Is PCA useful in this data?

Regarding my dataset, I think that PCA is not useful for this data for some reasons. PCA applies only for linear relationships but if we draw pair plot of this data, this shows many nonlinear relationships. (Pair plot is drawn below) Moreover in section (6), I chose two principal components but these components explains data only for about 48.45%.



8. Conclusion

To summarize my data, this data does not show multivariate normal distribution and there is not so much linear relationship between variables. I could check this by using histogram and probability plotting techniques. Also, I could find few significant correlation between living expense variables and other remaining variables.

By conducting PCA, I could get two principal components which are uncorrelated. Interpretation of those two components gave me quite meaningful results. However, these fail to explain big proportion of total variance in data.