FACTOR ANALYSIS

By
Suwanna Sayruamyat

Email: suwanna.s@ku.ac.th
Facebook: Suwanna Sayruamyat
Page: EatEcon
Website: www.eatecon.com

What is factor analysis?

- Factor analysis is an interdependence technique whose **primary purpose is to define the underlying structure among the variables in the analysis.**

- Variables are not classified as either dependent of independent. Instead, **the whole set of interdependent relationships among variables** is examined to define a set of common dimensions called **FACTORs**.

- Factor analysis is designed to **represent a wide range of attributes on a smaller number of new dimensions within data,** composite dimensions or factors with a minimum loss of information.

Aims of factor analysis

01 Data summarisation

02 Data reduction

03 Variable selection

- The goal of component analysis is to reduce a number of correlating variables to a smaller number of – usually uncorrelated – variables.

$$
\begin{aligned}
Y_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Y_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\
\vdots &= \vdots \\
Y_p &= a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p
\end{aligned}
$$

Where Xs are the original variables and Ys are the new variables (component scores). are component score coefficients. p is a number of components.

The sum of component variance is equal to the number of components.

Communalities

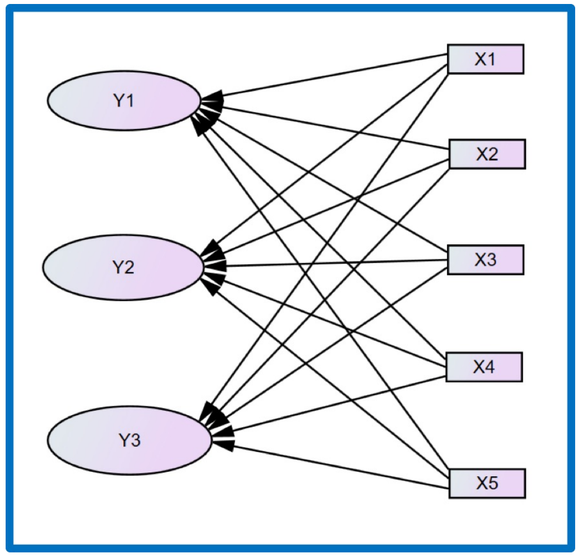
A communality is the extent to which **an item correlates with all other items.** Higher communalities are better. If communalities for a particular variable are low (between 0.0-0.4), then that variable may struggle to load significantly on any factor. In the table below, you should identify low values in the "Extraction" column. Low values indicate candidates for removal after you examine the pattern matrix.

Communalities		
	Initial	Extraction
Information about health risks caused by obesity, anorexia nervosa bulimia and other illesses liked to food	1	0.642
Information about foods related to health and beauty	1	0.617
Information about food containing anti ageing properties	1	0.625
Information about life style, food tourism and eating out	1	0.586
information about trends, consumption evolution, food fads, and underscoring ethnicity, cultural, social diversity of Italian polulation	1	0.475
Information about food safety issues caused by bacteria and other substances	1	0.492
Information about food regulations, affecting consumer choices and the food industry	1	0.51
Information about tradition, regional typical products and quality foods that are disappearing from the litalian market	1	0.566
Information about the production techniques used in the primary sector	1	0.761
Information about the food processing industry and innovations in terms of products and processes	1	0.737
Information about Italian and international cuisine, food culture and good living	1	0.556
Extraction Method: Principal Component Analysis.		

Component Matrix			
	Component		
	1	2	3
Information about health risks caused by obesity, anorexia nervosa bulimia and other illesses liked to food	0.403	0.626	0.296
Information about foods related to health and beauty	0.478	0.027	0.623
Information about food containing anti ageing properties	0.487	0.342	0.521
Information about life style, food tourism and eating out	0.539	-0.52	0.156
information about trends, consumption evolution, food fads, and underscoring ethnicity, cultural, social diversity of Italian polulation	0.546	-0.39	0.156
Information about food safety issues caused by bacteria and other substances	0.528	0.461	-0.033
Information about food regulations, affecting consumer choices and the food industry	0.577	0.306	-0.288
Information about tradition, regional typical products and quality foods that			

Component analysis VS Factor analysis

Component analysis

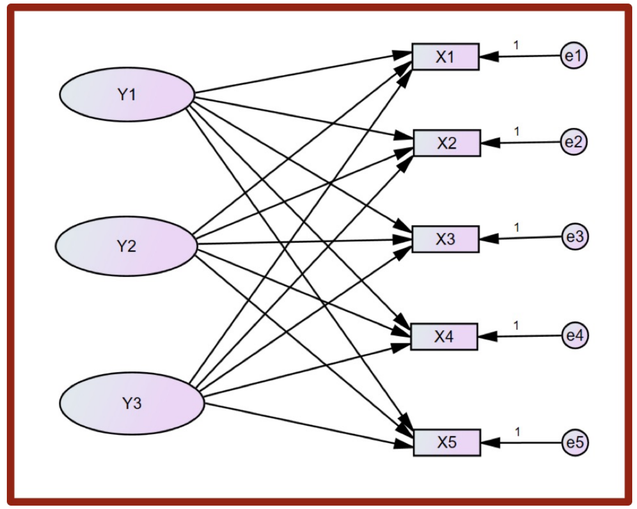


$$Y_i = \sum_{j=1}^{p} a_{ij} x_j for i = 1,2,\ldots,p$$

Formative

- Direction of causality is from measure to construct
- No reason to expect the measures are correlated
- Indicators are not interchangeable

Factor analysis



$$X_j = \sum_{j=1}^{p} b_{ji} Y_i for j = 1,2,\ldots,p$$

$$X_j = \sum_{j=1}^{p} \lambda_{ji} F_i + \lambda F_{j.spec} + e_j for j = 1,2,\ldots,p$$

Reflective

- Direction of causality is from construct to measure
- Measures expected to be correlated
- Indicators are interchangeable

If you want **summarising** a number of correlating variables in a few new variable with smallest possible loss of information, the **component analysis** is the answer.

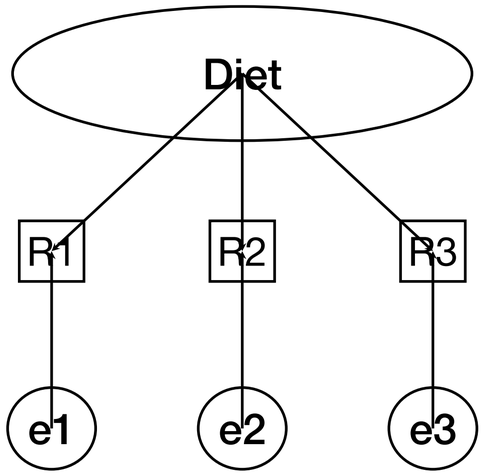
If you want **explaining the correlations** in a data set in form of factors, the **factor analysis** is the answers.

However, component analysis is less complicated and usually give the same results as exploratory factor analysis. Thus, **most component analysis and EFA both go under the name of factor analysis (Blunch, 2013).**

Formative vs. Reflective

Reflective

R1. I eat healthy food.
R2. I do not each much junk food.
R3. I have a balanced diet.



Formative

F1. I have a balanced diet.
F2. I exercise regularly.
F3. I get sufficient sleep each night.

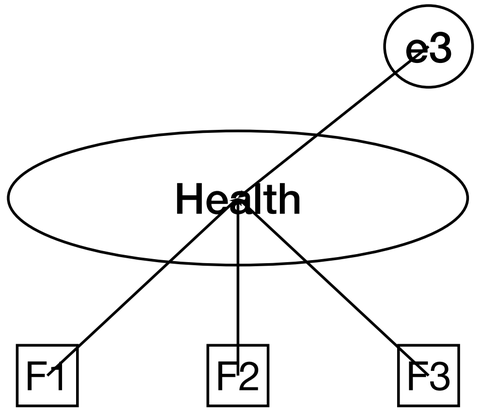


Photo by http://statwiki.kolobkreations.com/index.php?title=Exploratory_Factor_Analysis#Communalities

Types of factor analysis

Exploratory factor analysis (EFA)

- Summarising data by grouping correlated variables.
- Investigating sets of measured variables related to theoretical constructs.
- Preliminary exploration of data (**Data-driven**)

Confirmatory factor analysis (CFA)

- Testing generalisation of factor structure to new data.
- Making use of only the measurement model component of the general SEM.
- It should be based on theory and/or the results of EFA and other psychometric tests.
- Test of theory against data(**Theory-driven**)

Exploratory factor analysis (EFA)

- Also called “unrestricted” factor analysis.
- Finds factor loadings which best reproduce correlations between observed variables.
- n factors = n of observed variables.
- All variables related to all factors.
- Retain <n factors which ‘explain’ satisfactory amount of observed variance.
- ‘Meaning’ of factors determined by pattern of loadings.
- No unique solution where >1 factor, rotation used to clarify what each factor measures.

Assumptions

- Based on linear relationships.
- The data collected are interval scaled.
- Multicollinearity in the data is desirable because the aim is to identify interrelated set of variables.
- The data should not be a variable that only correlate with itself and no correlation exists with any other variables.
- Data is not an identity matrix.

Sample size

1. Minimum number of variable for FA is 5 observations per variable (1:5)

   Ex. 20 variables should have $\geq$ 100 observations.

2. Ideal condition ratio is 1:20.

   Ex. 20 variables ideally should > 400 observations.

3. The sample must have more observations than variables.

4. The minimum absolute sample size should be 50 observations.

Source: Hair et al. (2014)

Reliability

The reliability of k items is **Cronbach's alpha** => $\hat{\alpha}_{std} = \frac{k\bar{r}}{1+(k-1)\bar{r}}$

Where
- k = no. of items
- std =standardised
- $\bar{r}$ = the average correlation

SPSS

1. Click Analyse > Scale > Reliability Analysis …
2. Select the items in the left block adding to the right block.
3. Click 'Statistic' > continue
4. In the model pane choose 'Alpha'
5. OK

Rotation options

Orthogonal

- Maintains independence of factors
- More commonly seen
- Usually at least one option
- Method: varimax, quartimax, equamax, parsimax, etc.

Oblique

- Allows dependence of factors
- Make distinctions sharper (loadings closer to 0's and 1's.
- Can be harder to interpret once you lose independence of factors
- Method: promax, oblimin, etc.

Uniqueness

- Uniqueness for each item describes **the proportion of the item described by the factor model.**
- Recall an R-squared: Proportion of variance in Y explained by X.
- 1-Uniqueness: proportion of the variance in X_k explained by F1, F2, etc.
- Uniqueness: represents what is left over that is not explained by factors
- A GOOD item has a LOW uniqueness

Convergent validity

Convergent validity means that the variables within a single factor are highly correlated. This is evident by the factor loadings. Sufficient/significant loadings depend on the sample size of your dataset. The table below outlines the thresholds for sufficient/significant factor loadings. Generally, the smaller the sample size, the higher the required loading. We can see that in the pattern matrix above, we would need a sample size of 60-70 at a minimum to achieve significant loadings for variables loyalty1 and loyalty7. Regardless of sample size, it is best to have loadings greater than 0.500 and averaging out to greater than 0.700 for each factor.

Sample size	Sufficient factor loading
50	0.75
60	0.70
70	0.65
85	0.60
100	0.55
120	0.50
150	0.45
200	0.40
250	0.35
350	0.30

http://statwiki.kolobkreations.com/index.php?title=Exploratory_Factor_Analysis#Communalities

Steps in EFA

1. Collect and explore data: choose relevant variables
2. Determine the number of factors
3. Estimate the model using predefined number of factors
4. Rotate and interpret
5. Decide if changes need to be made (e.g. drop item(s), include item(s)) and repeat step 3 - 4
6. Construct scales and use in further analysis

Data for practice

For excel : https://www.eatecon.com/courses/training/?tab=tab-overview

For SPSS

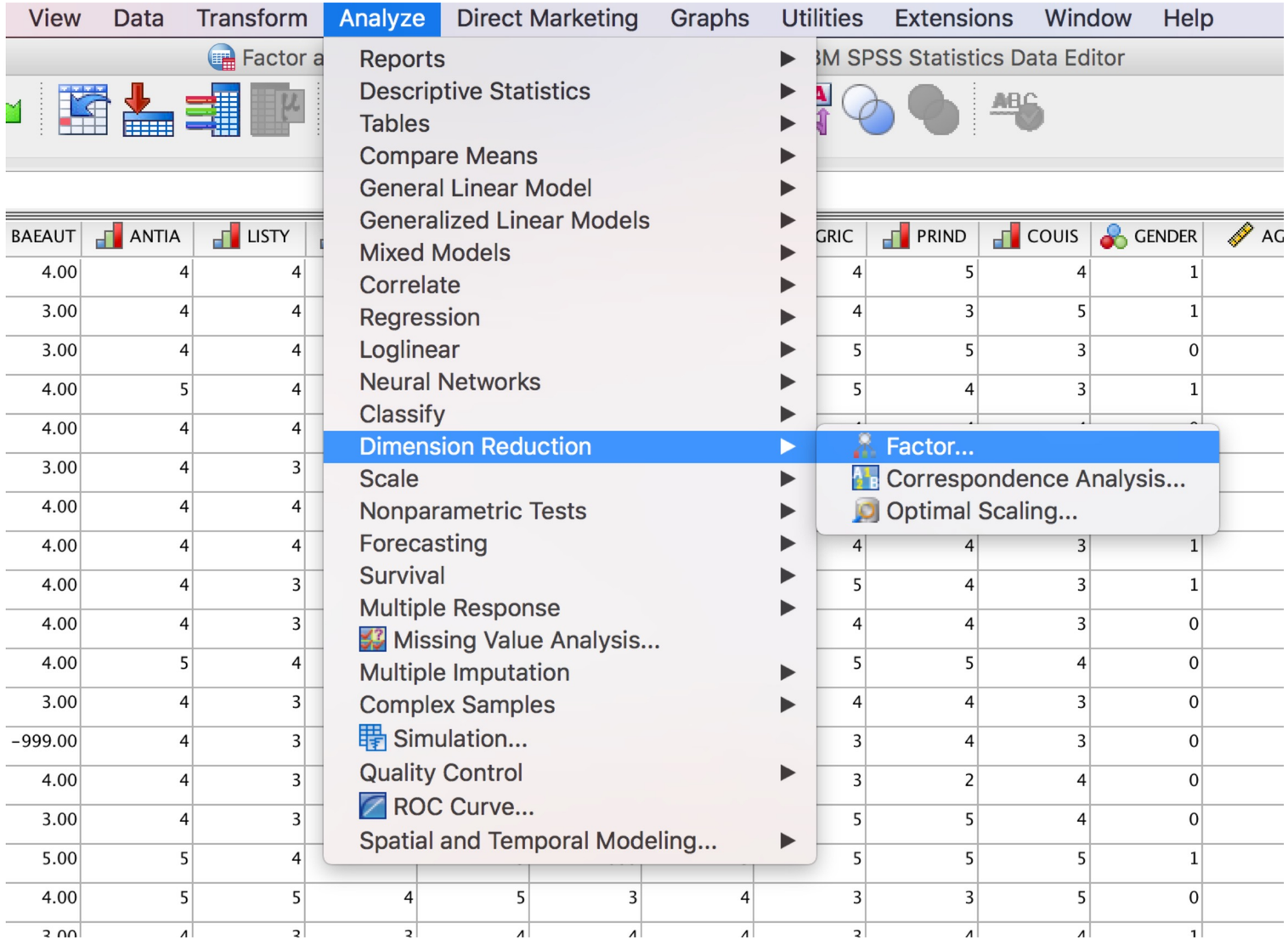
Google classroom

Class code: 7dl3bgj

Practice

Step 1:

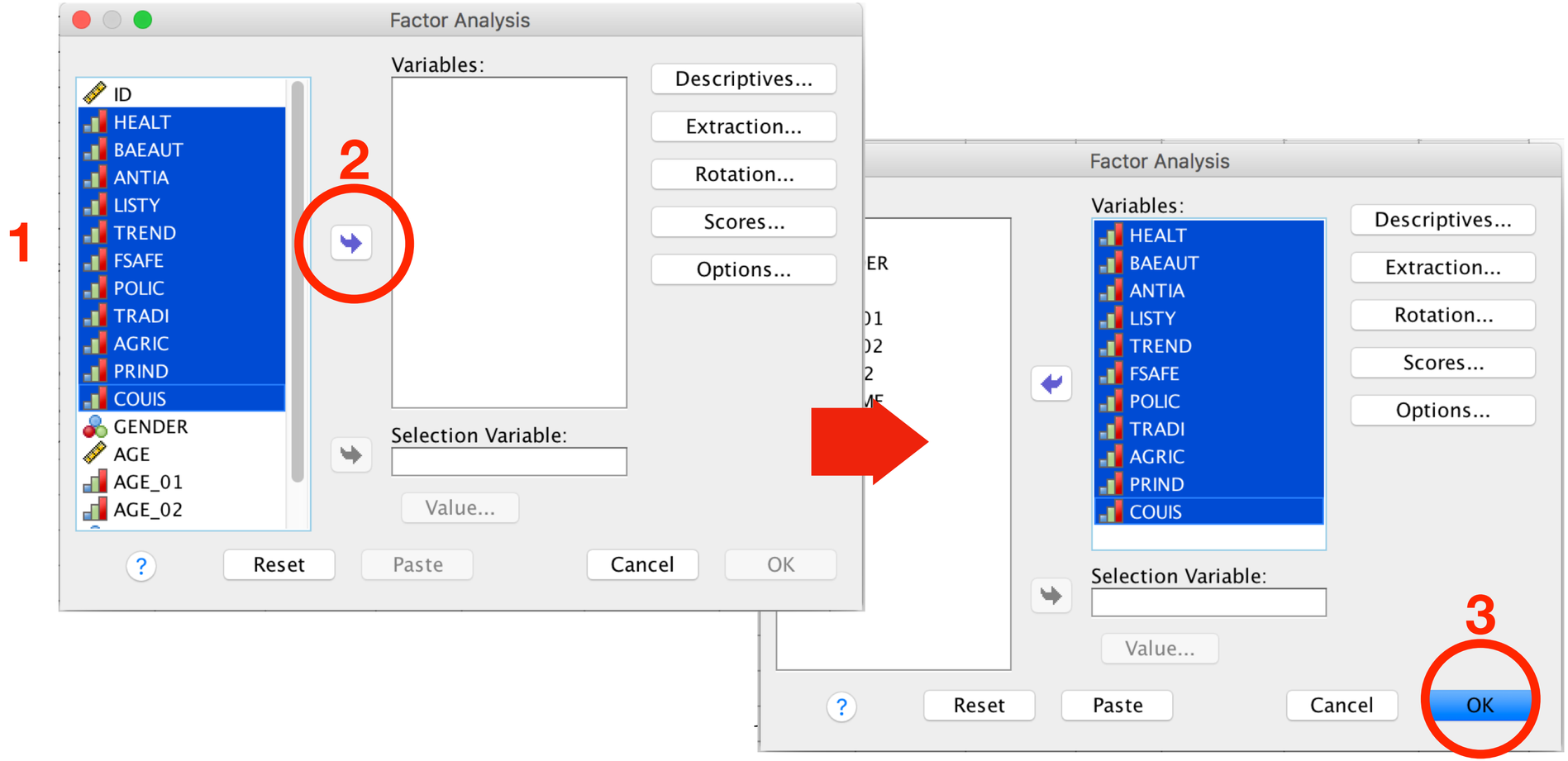
- Download the file Factor analysis - **Food inf preferences.sav**
- To perform a factor analysis: click the **Analyse > Dimension Reduction > Factor**



Practice

Step 2:

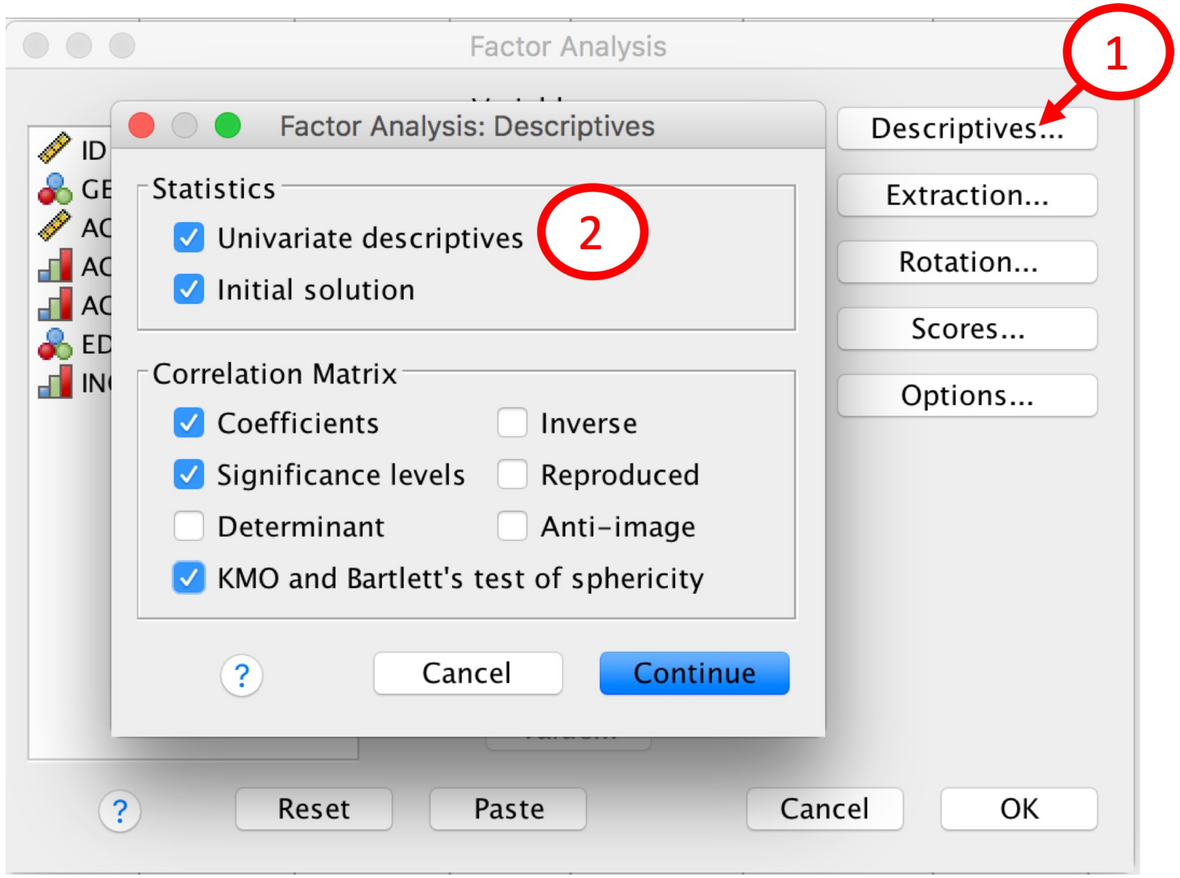
- Select 11 attitudinal variables (HEALT, BAEAUT, ANTIA, LISTA, TREND, FSAFE, POLIC, TRADI, AGRIC, PRIND, and COUIS) from the left hand side dialogue box and click the blue arrow



Practice

Step 3:

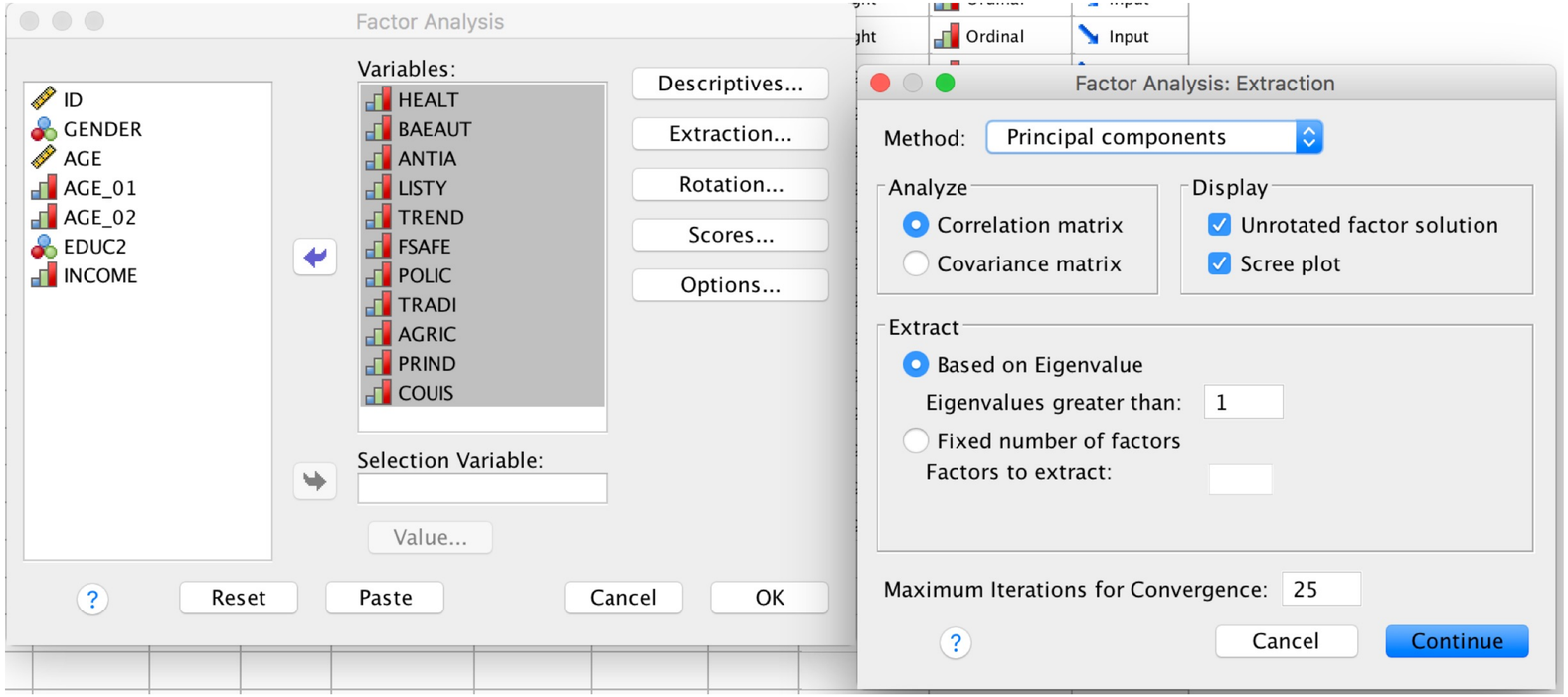
- Click on Descriptives… to open the Factor analysis Descriptives box
- Select:
  - ‘Univariate descriptive’,
  - ‘Initial solutions’
  - ‘KMO and Barlett’s test of sphericity’
- Click on Continue;



Practice

Step 4:

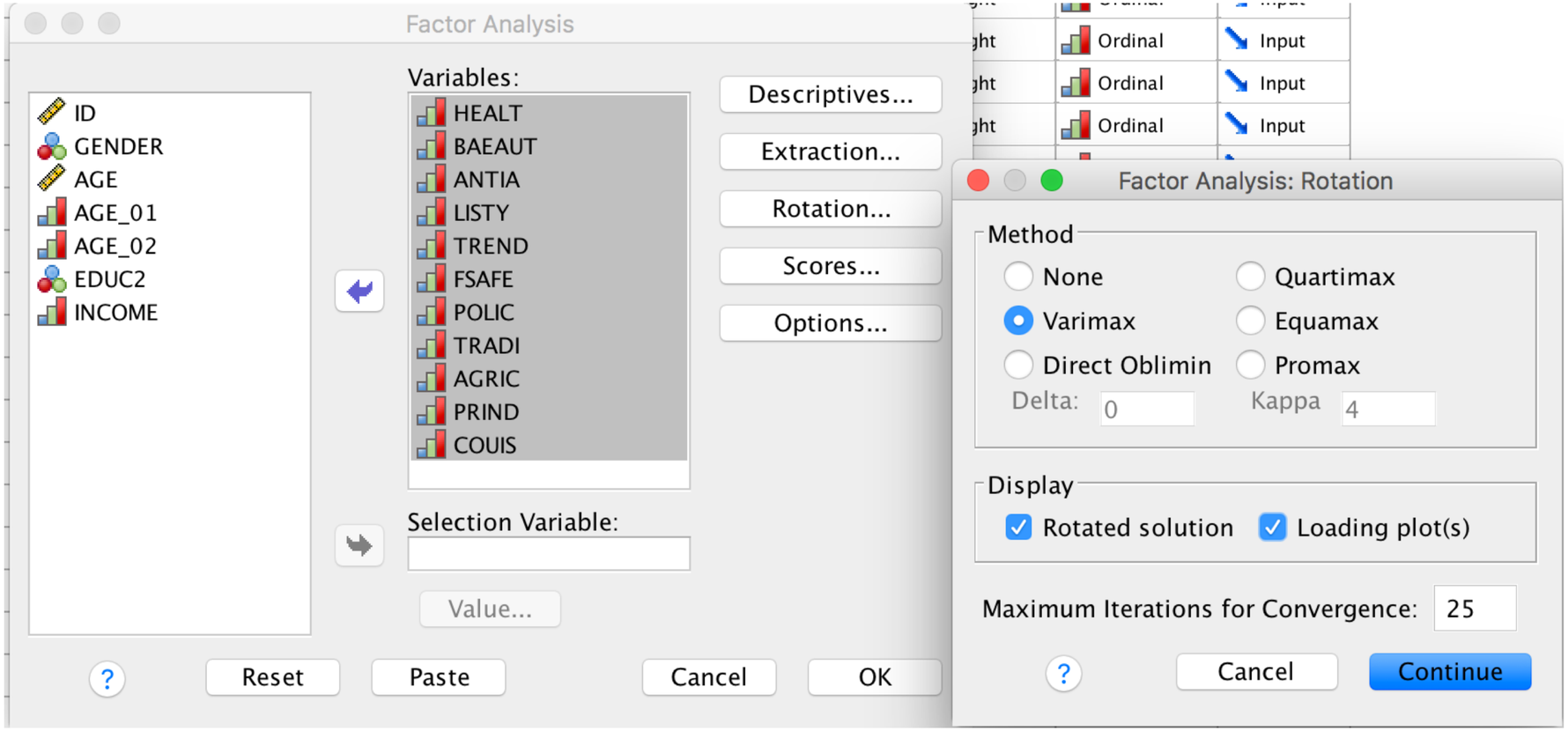
- Click on **Extraction** to open 'Factor analysis Extraction' box
- Select '**Principal components**' method, '**Correlation matrix'**, '**Unrotated factor solutions'** and '**Scree plot'**
- Click on **Continue**;



Practice

Step 5:

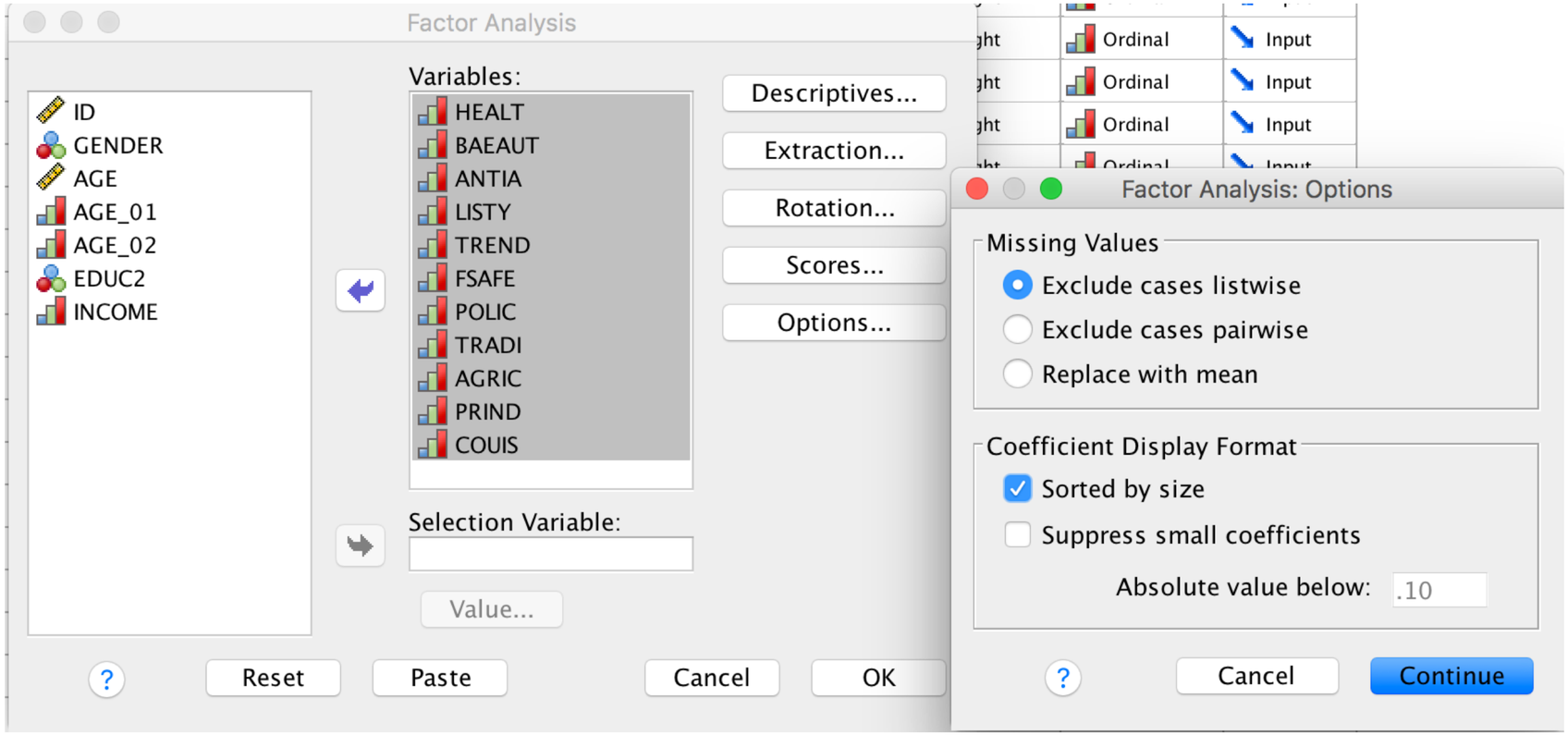
- Click on Rotation to open the ‘Factor Analysis Rotation’ box
- Select Varimax, ‘Rotated solution’ and ‘Loading plot(s)’
- Click on Continue;



Practice

Step 6:

- Click on Options to open the ‘Factor Analysis Options’ box
- Select ‘Sorted by size’
- Click on Continue;
- Next Click on OK.



Test for an identity matrix

Note: Correlation matrix => **highly correlated** variables indicate that factor analysis may be an **appropriate** multivariate statistical technique to explore these variables.

Correlation Matrix

		Information about health risks caused by obesity, anorexia nervosa bulimia and other illesses liked to food	Information about foods related to health and beauty	Information about food containing anti ageing properties	Information about life style, food tourism and eating out	information about trends, consumption evolution, food fads, and underscoring ethnicity, cultural, social diversity of Italian polulation	Information about food safety issues caused by bacteria and other substances	Information about food regulations, affecting consumer choices and the food industry	Information about tradition, regional typical products and quality foods that are disappearing from the litalian market	Information about the production techniques used in the primary sector	Information about the food processing industry and innovations in terms of products and processes	Information about Italian and international cuisine, food culture and good living
Correlation	Information about health risks caused by obesity, anorexia nervosa bulimia and other illesses liked to food	1.000	.262	.384	.007	.028	.401	.241	.105	.170	.195	.061
	Information about foods related to health and beauty	.262	1.000	.429	.217	.286	.149	.134	.210	.117	.138	.281
	Information about food containing anti ageing properties	.384	.429	1.000	.152	.173	.255	.194	.143	.187	.203	.150
	Information about life style, food tourism and eating out	.007	.217	.152	1.000	.394	.111	.115	.417	.240	.203	.415
	information about trends, consumption evolution, food fads, and underscoring ethnicity, cultural, social diversity of Italian polulation	.028	.286	.173	.394	1.000	.123	.218	.336	.276	.203	.326
	Information about food safety issues caused by bacteria and other substances	.401	.149	.255	.111	.123	1.000	.376	.212	.289	.310	.142
	Information about food regulations, affecting consumer choices and the food industry	.241	.134	.194	.115	.218	.376	1.000	.302	.402	.400	.087

Appropriateness of data (adequacy)

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.775
Bartlett's Test of Sphericity	Approx. Chi-Square	2241.220
	df	55
	Sig.	.000

Note:

- Kaiser-Meyer-Olkin value should be large (>.5).
- Bartlett's test should be less (p-value <.05).
- Null hypothesis: the correlation matrix of population under investigation is not an identity matrix.

KMO statistics

- Marvelous: .90s
- Meritorious: .80s
- Middling: .70s
- Mediocre: .60s
- Miserable: .50s
- Unacceptable: <.50

Bartlett's Test of Sphericity

A significant result (Sig. < 0.05) indicates matrix is not an identity matrix; i.e., the variables do relate to one another enough to run a meaningful EFA.

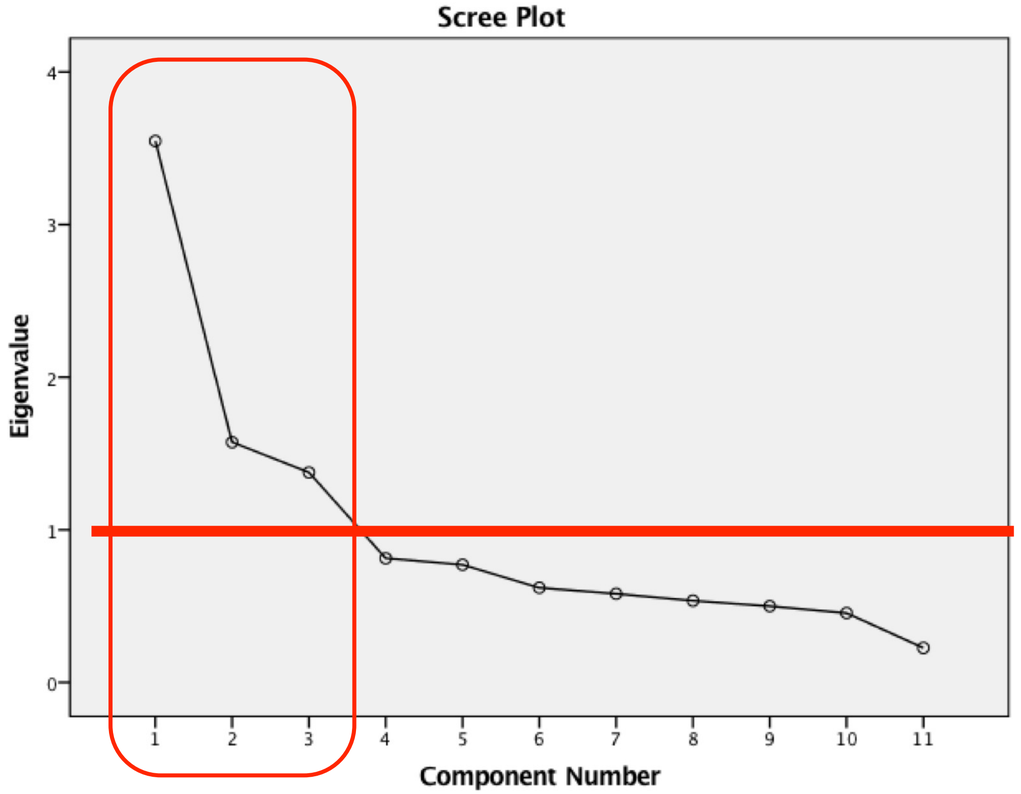
How many factors should be selected?

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.547	32.248	32.248	3.547	32.248	32.248	2.340	21.270	21.270
2	1.574	14.309	46.557	1.574	14.309	46.557	2.323	21.118	42.389
3	1.376	12.512	59.069	1.376	12.512	59.069	1.835	16.680	59.069
4	.813	7.395	66.464						
5	.771	7.013	73.477						
6	.620	5.638	79.114						
7	.581	5.280	84.395						
8	.536	4.868	89.263						
9	.500	4.543	93.807						
10	.455	4.134	97.941						
11	.226	2.059	100.000						

Extraction Method: Principal Component Analysis.

$$Initial\ Eigenvalue = \frac{total\ variance}{total\ no.\ of\ component}$$



Factor1 accounts for 32.57% of the total variance (3.594/11)
Factor2 accounts for 14.42% of the total variance (1.574/11)
Factor3 accounts for 12.51% of the total variance (1.376/11)

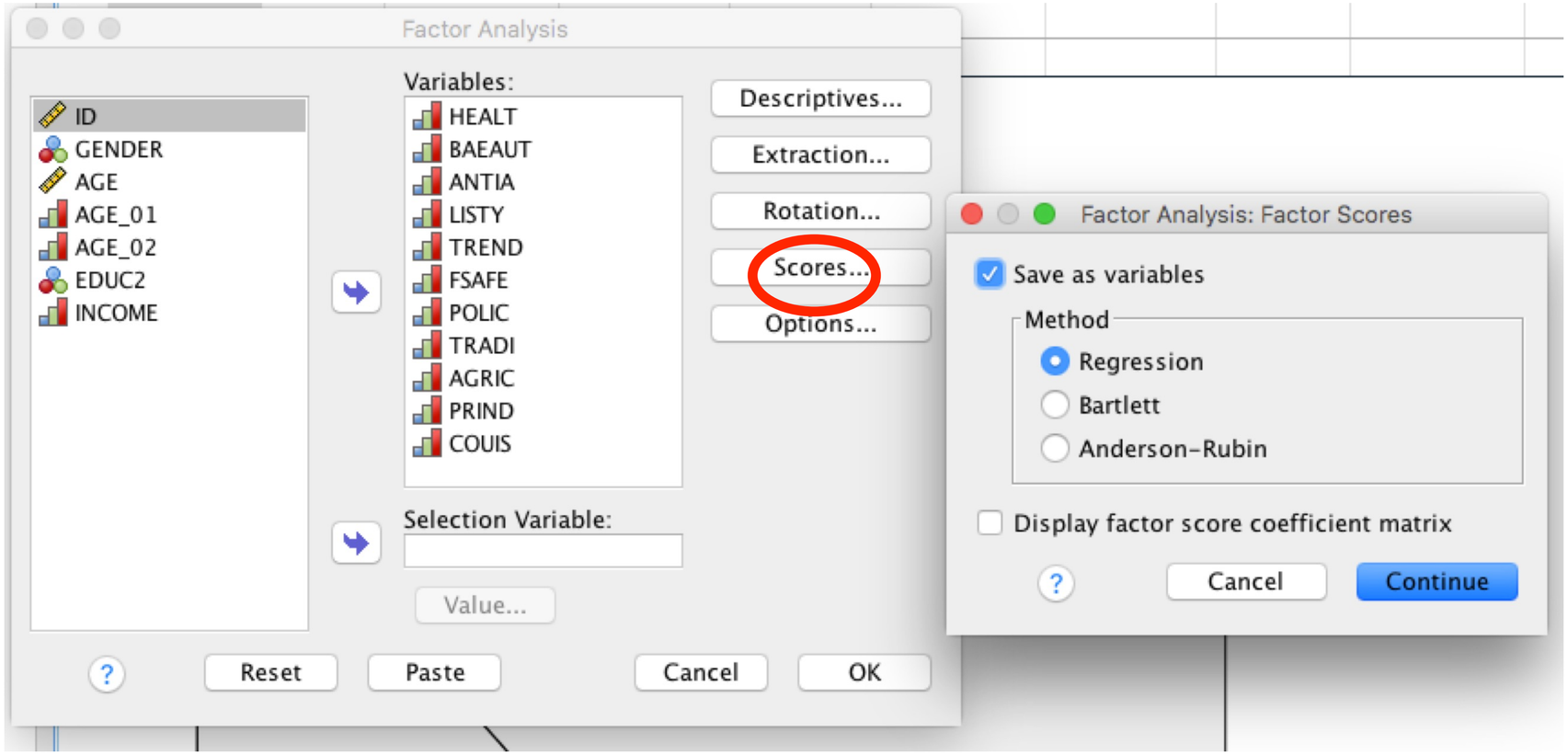
Therefore, first tree factors explain 59.7% of total variance, which mean new variables contain the information of original variables by 59.75%.

Rotated Component Matrix

		Component		
		1	2	3
Information about the food processing industry and innovations in terms of products and processes	PRIND	0.825	0.223	0.038
Information about the production techniques used in the primary sector	AGRIC	0.82	0.286	-0.005
Information about food regulations, affecting consumer choices and the food industry	POLIC	0.667	0.057	0.22
Information about food safety issues caused by bacteria and other substances	FSAFE	0.52	-0.025	0.465
Information about life style, food tourism and eating out	LISTY	0.071	0.753	0.027
Information about Italian and international cuisine, food culture and good living	COUIS	0.104	0.735	0.083
information about trends, consumption evolution, food fads, and underscoring ethnicity, cultural,	TREND	0.119	0.663	0.116

How to save new variables

1. Redo step1 – 6
2. Step 7: Click on Scores to open the ‘Factor Analysis: Factor Scores’ box
3. Select ‘Save as variables’ and Regression method
4. Click on Continue and OK



Limitations of EFA

- Inductive, a theoretical (Data->Theory)
- Subjective judgement & heuristic rules
- We usually have a theory about how indicators are related to particular latent variables (Theory-> Data)
- Be explicit and test this measurement theory against sample data

References

1.http://statwiki.kolobkreations.com/index.php?title=Exploratory_Factor_Analysis#Communalities