การวิเคราะห์องค์ประกอบ

(Factor analysis)

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

BA603

เป้าหมายของการวิเคราะห์องค์ประกอบ

Data summarisations

Data reduction

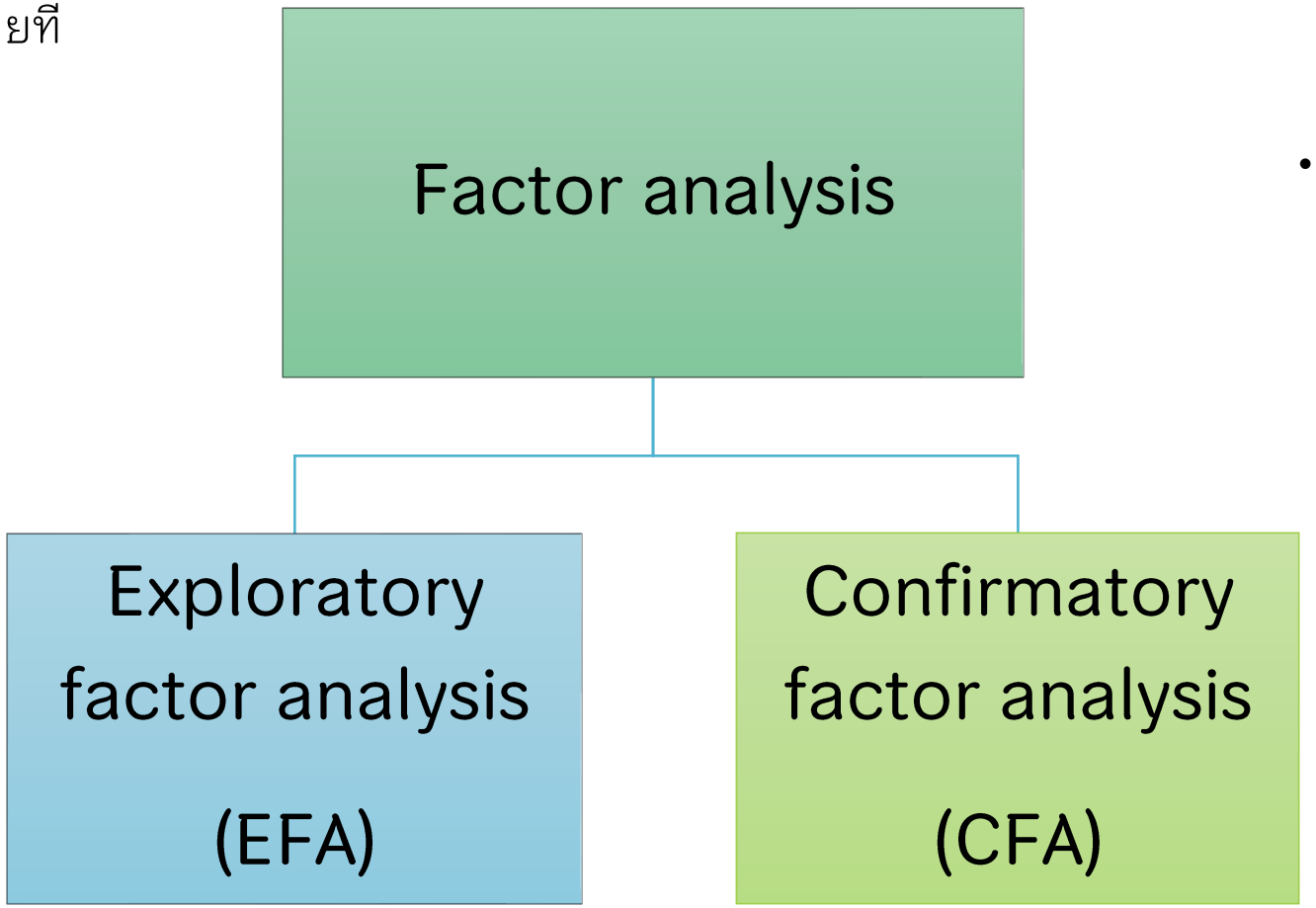
Variable selection

1. การวิเคราะห์องค์ประกอบ (Factor analysis) เป็นเทคนิคในการค้นหาตัวประกอบต่าง ๆ (factors) จากชุดตัวแปรที่มีองค์ประกอบร่วมกัน (สัมพันธ์กัน)
   - เป้าหมายหลักคือจุดประสงค์หลักเพื่อกำหนดโครงสร้างพื้นฐานของตัวแปรในการวิเคราะห์
2. ตัวแปรไม่ได้ถูกระบุประเภทว่าเป็นตัวแปรตามหรือตัวแปรอิสระ แต่จะตรวจสอบชุดความสัมพันธ์ที่พึ่งพาซึ่งกันและกันทั้งหมดระหว่างตัวแปร เพื่อกำหนดชุดของมิติร่วมที่เรียกว่า ปัจจัย (FACTORS)
3. การวิเคราะห์องค์ประกอบได้รับการออกแบบมาเพื่อแสดงคุณลักษณะที่หลากหลายในจำนวนมิติใหม่ที่มีจำนวนน้อยลงโดยมีการสูญเสียข้อมูลน้อยที่สุด

ประเภทของการวิเคราะห์องค์ประกอบ

EFA: ต้องการจัดกลุ่มตัวแปร โดยที่ยังไม่รู้มาก่อนว่าตัวแปรใดอยู่ภายใต้องค์ประกอบใด

- สำรวจเพื่อให้รู้ว่าตัวแปรใดอยู่ด้วยกันบ้าง



CFA: ต้องการตรวจสอบตัวแปรที่อยู่ในแต่ละกลุ่มว่ามีน้ำหนักเพียงพอที่จะอยู่ในกลุ่มนั้น จริงหรือไม่

- มีกรอบแนวคิดอยู่แล้ว ต้องการยืนยันว่าตัวแปรเหล่านั้นอยู่ในกลุ่มจริง
- ใช้ CFA เพื่อทดสอบว่าโมเดลที่กำหนดไว้เหมาะสมกับข้อมูลที่มีมากน้อยเพียงใด

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

- 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)

การวิเคราะห์องค์ประกอบ

- Also called “unrestricted” factor analysis.
- ค้นหาค่าความสัมพันธ์ของปัจจัย (factor loadings) ที่สร้างความสัมพันธ์ระหว่างตัวแปรที่สังเกตได้ให้ดีที่สุด
- จำนวนปัจจัย (n factors) = จำนวนตัวแปรที่สังเกตได้ (n of observed variables)
- ตัวแปรทั้งหมดมีความสัมพันธ์กับทุกปัจจัย (factors).
- เก็บรักษาจำนวนปัจจัย < n ที่ "อธิบาย" ปริมาณความแปรปรวนที่สังเกตได้ในระดับที่น่าพอใจ
- "ความหมาย" ของปัจจัยถูกกำหนดโดยรูปแบบของค่าความสัมพันธ์ของปัจจัย (pattern of loadings).
- No unique solution where >1 factor, rotation used to clarify what each factor measures.

สมมติฐานการวิเคราะห์

1. ความสัมพันธ์ระหว่าง Factor กับ variable เป็นเชิงเส้นตรง (linear relationship)
2. ข้อมูลที่ใช้ควรเป็น interval scale
3. Factor และ error เป็นอิสระต่อกัน
4. จำนวนข้อมูลที่นำมาวิเคราะห์ต้องมากกว่าจำนวนตัวแปร
5. Multicollinearity in the data is desirable because the aim is to identify interrelated set of variables.
6. Data is not an identity matrix. (ตอบเหมือนๆ กัน)

ขนาดตัวอย่างที่เหมาะสมในการวิเคราะห์

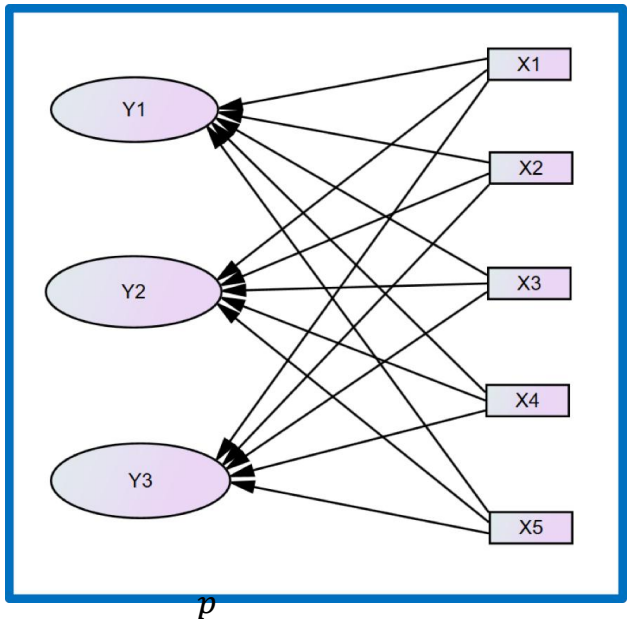
1. อัตราส่วนขั้นต่ำของตัวแปรสำหรับ EFA คือ 1:5 (1 ตัวแปร ต่อจำนวนตัวอย่าง 5 observations
   - Ex. 20 ตัวแปร ควรมีตัวอย่าง ≥ 100 observations
2. Ideal condition ratio is 1:20.
3. จำนวนตัวอย่างต้องมากกว่าจำนวนตัวแปรที่วิเคราะห์
4. ขนาดตัวอย่างไม่ควรน้อยกว่า 50 ตัวอย่าง

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

Source: Hair et al. (2014)

Component analysis VS Factor analysis

Principle component analysis

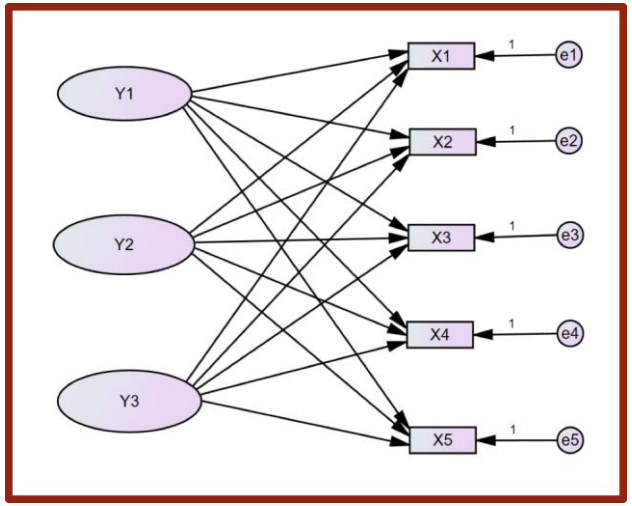


$$Y_i = \sum_{j=1}^{p} a_{ij}x_j for i = 1,2,\ldots,p$$

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

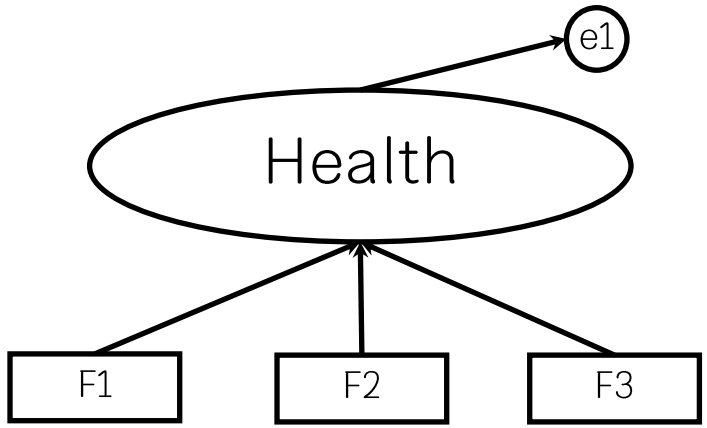
- Direction of causality is from construct to measure
- Measures expected to be correlated
- Indicators are interchangeable

- 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

Formative

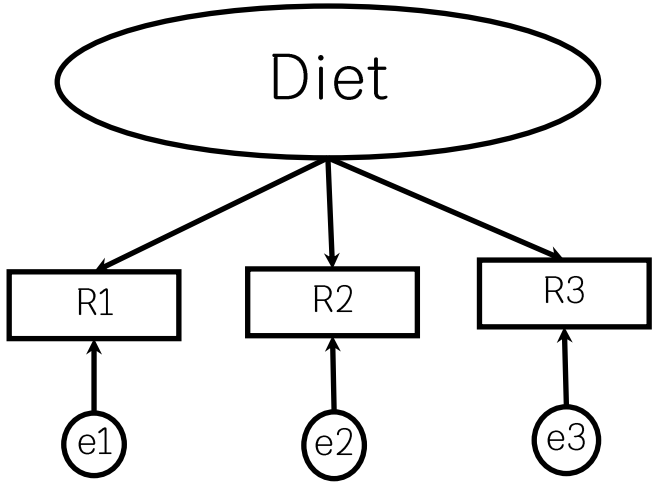
F1. I have a balanced diet.
F2. I exercise regularly.
F3. I get sufficient sleep each night.



Principle component analysis

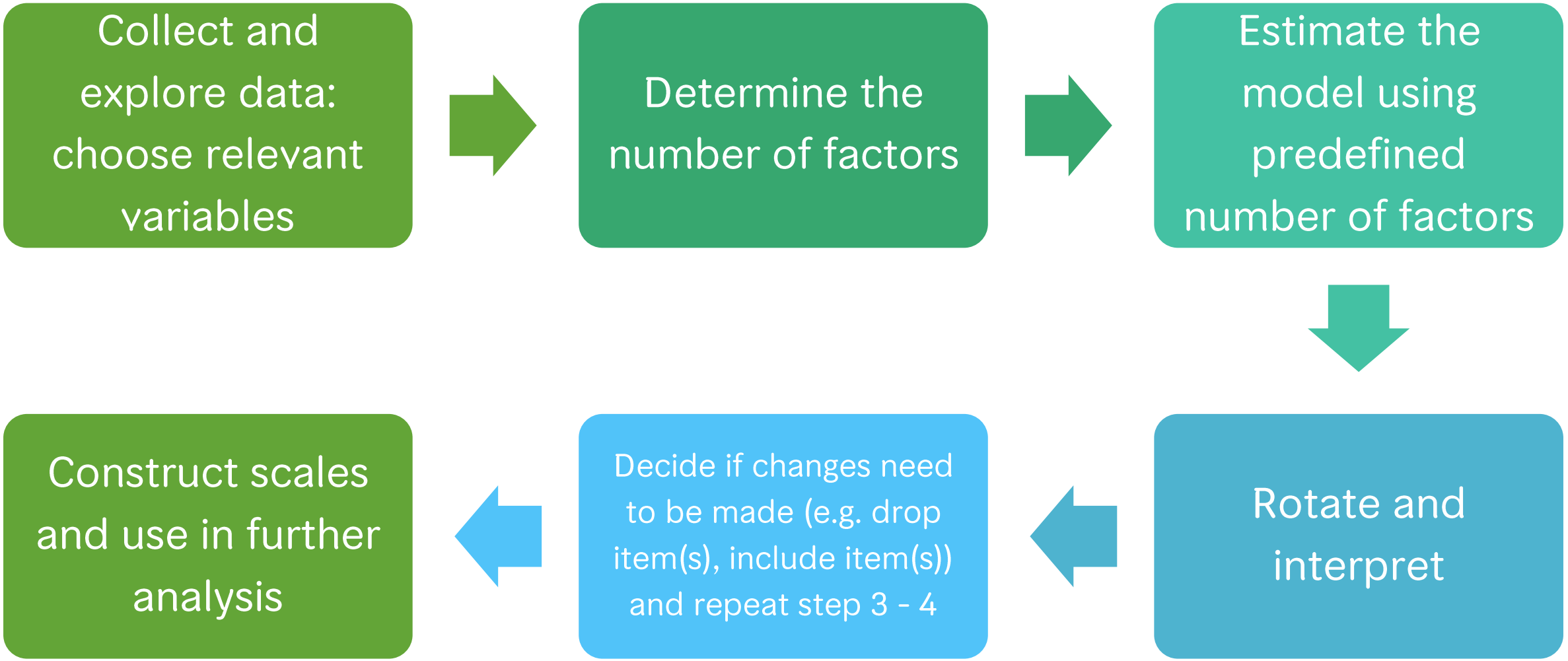
Reflective

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



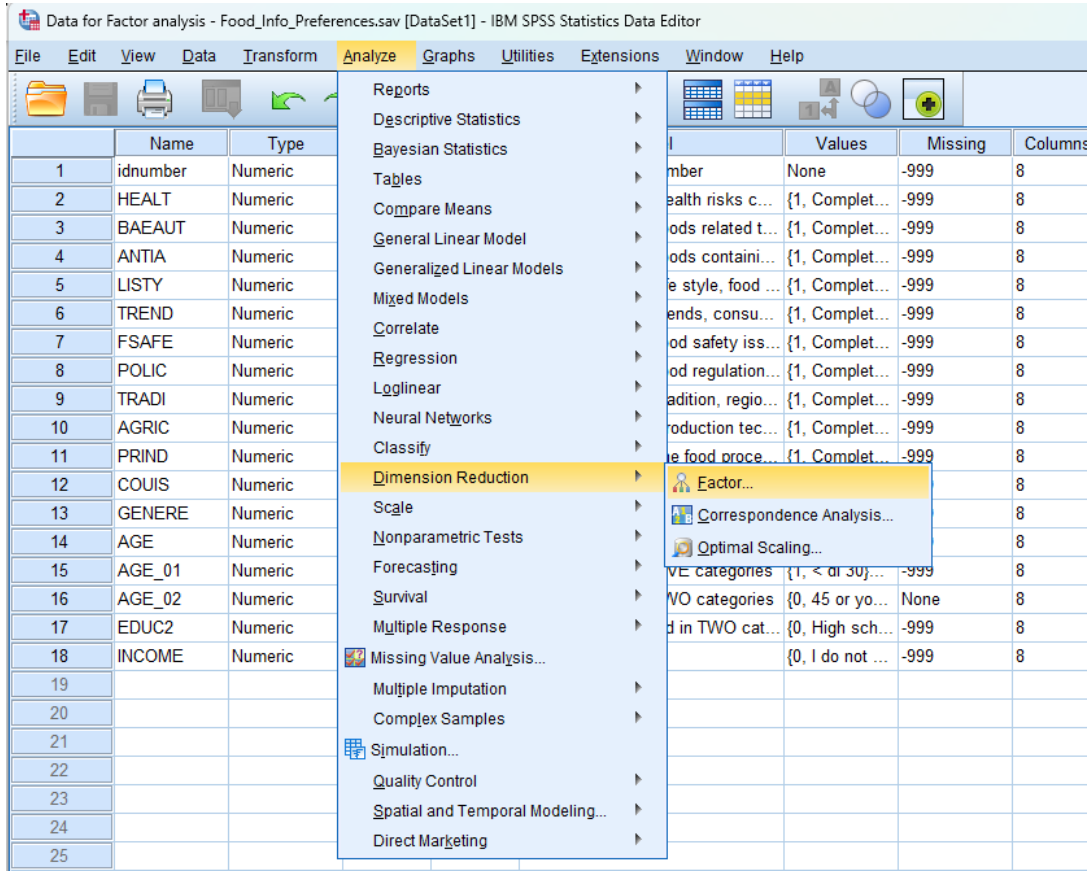
Factor analysis

Steps in EFA

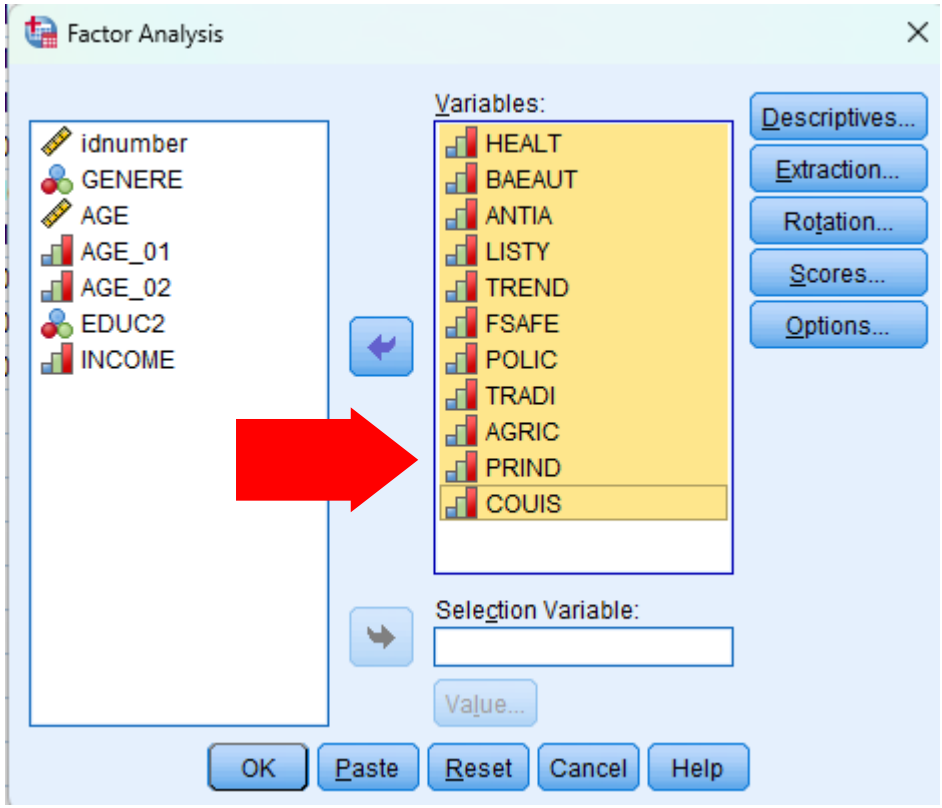

Collect and explore data: choose relevant variables
Determine the number of factors
Estimate the model using predefined number of factors
Rotate and interpret
Decide if changes need to be made (e.g. drop item(s), include item(s)) and repeat step 3 - 4
Construct scales and use in further analysis

1. Download the file Factor analysis – **Data for factor analysis**

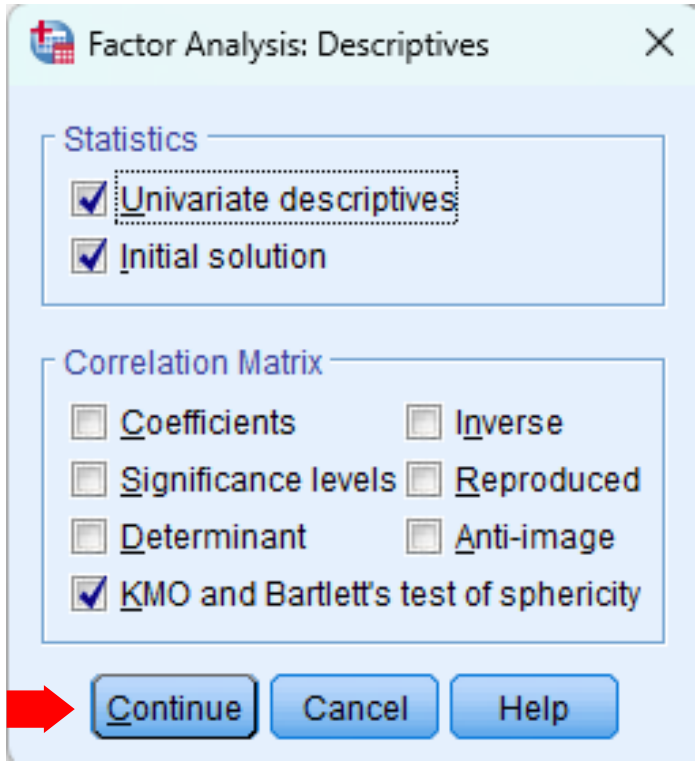
2. click the Analyse > Dimension Reduction > Factor



เลือกตัวแปรที่ต้องการใส่ในช่อง variables



Factor analysis: Descriptives



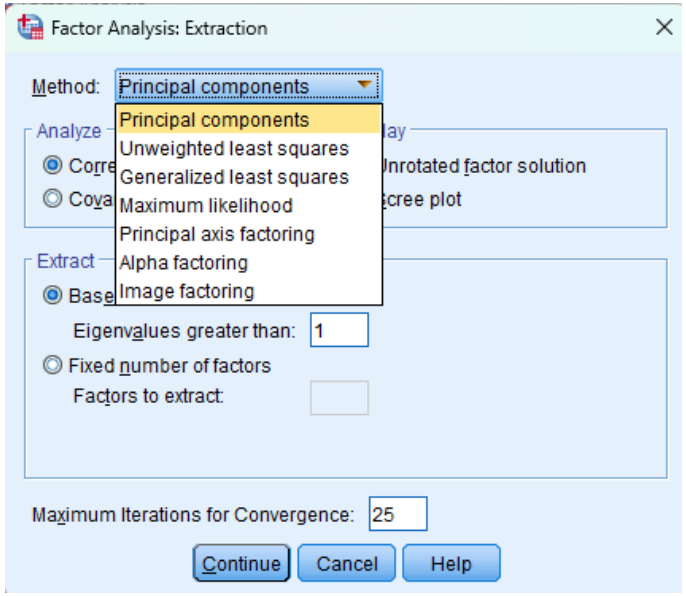
Statistics

- Univariate descriptive แสดงจำนวนข้อมูล , ค่าเฉลี่ย และค่าเบี่ยงเบนมาตรฐานของตัวแปรแต่ละตัว
- Initial solution แสดงค่า initial communalities, eigenvalue และ percentage of variance explained

Correlation Matrix

- Coefficients แสดงค่าเมทริกซ์สัมประสิทธิ์สหสัมพันธ์ของตัวแปรทุกคู่
- Significance levels แสดงค่า one-tailed significance level ของการทดสอบค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปรแต่ละคู่
- Determinant แสดงค่า determinant ของเมทริกซ์สัมประสิทธิ์สหสัมพันธ์
- KMO and Bartlett’s test of sphericity แสดงค่า KMO และBartlett’s test
  - KMO (Kaiser-Meyer-Olkin) เป็นค่าที่ใช้วัดความเหมาะสมของข้อมูลตัวอย่างที่จะนำมาวิเคราะห์โดยเทคนิค Factor Analysis

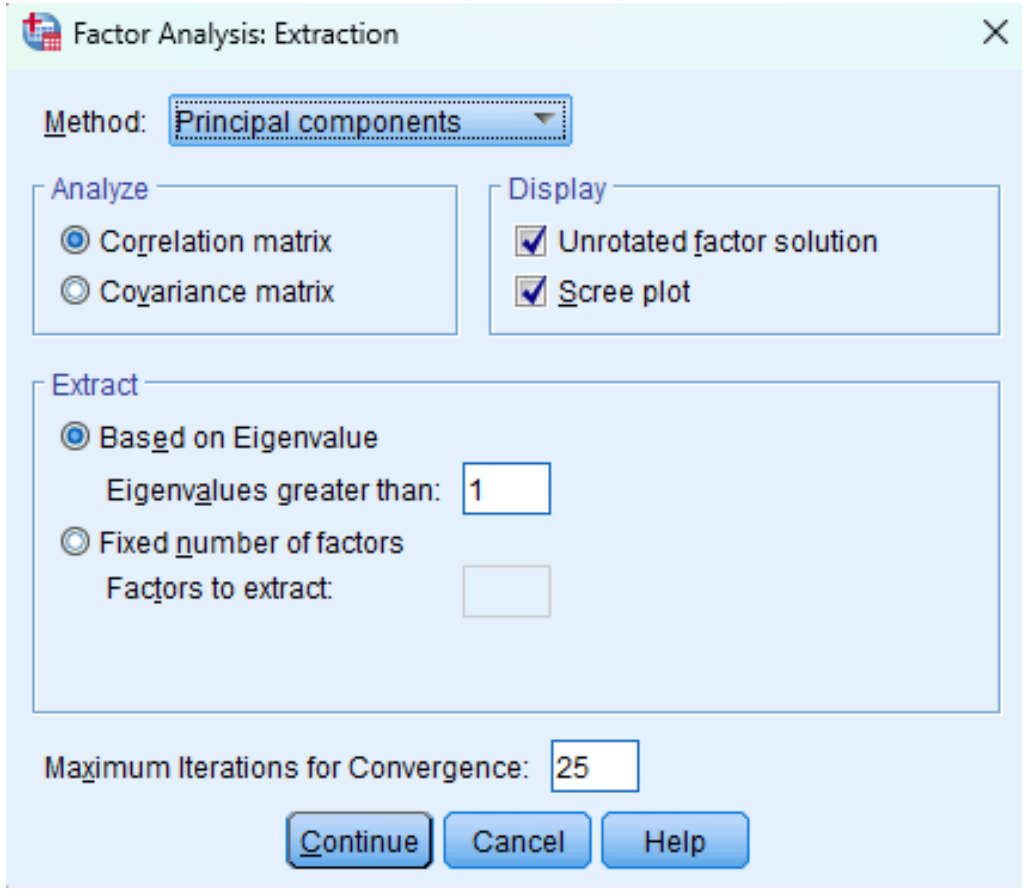
Factor analysis: Extraction



Method เลือกเทคนิคการวิเคราะห์ปัจจัย แบ่งออกเป็น 2 วิธีหลัก คือ

1. **Principal Component Analysis (PCA)** เป็นวิธีแบบ formative ความนิยมมากที่สุด
2. **Common Factor Analysis (CFA)** เป็นเทคนิคที่มีวัตถุประสงค์เหมือนเทคนิค PCA คือ จะสร้าง Factor เพื่อลดจำนวนตัวแปร แต่หลักเกณฑ์ของ CFA จะพยายามทำให้ค่าแปรปรวนเฉพาะส่วนของ common factor มากที่สุด โดยไม่พิจารณาถึงค่า Unique Factor เทคนิค CFA มีเทคนิคย่อย 6 เทคนิคดังนี้
   1) Unweighted Least Square เป็นเทคนิคที่ต้องกำหนดจำนวน factor ไว้แน่นอนก่อน แล้วหา Factor pattern matrix ที่ทำให้ผลบวกกำลังสองของระยะห่างระหว่างเมตริกซ์สัมประสิทธิ์สหสัมพันธ์ที่คำนวณได้จากข้อมูล กับเมตริกซ์สัมประสิทธิ์สหสัมพันธ์ที่สร้างขึ้นใหม่ให้มีค่าน้อยที่สุด
   2) Generalized Least Square มีหลักเกณฑ์เหมือนวิธี Unweighted Least Square แต่จะมีการถ่วงน้ำหนักค่าสัมประสิทธิ์สหสัมพันธ์ ด้วยค่าผกผันของ Uniques ของตัวแปรนั้น นั่นคือจะให้น้ำหนักแก่ตัวแปรที่มีค่าUnique สูงน้อยกว่าตัวแปรที่มีค่า unique ต่ำ
   3) Maximum Likelihood Method วิธีนี้กำหนด factor โดยการประมาณค่าพารามิเตอร์ที่ทำให้ เมตริกซ์สัมประสิทธิ์สหสัมพันธ์ที่คำนวณได้ มีค่าใกล้กับเมตริกซ์ที่ได้จากข้อมูล โดยมีเงื่อนไขว่า ข้อมูลตัวอย่างนั้น (ตัวแปร) ต้องมีการแจกแจงแบบ Multivariate Nomal
   4) Alpha Method
   5) Image Factoring

Factor analysis: Extraction



Display

- ✓ Unrotate factor solution เมื่อต้องการให้แสดงผลลัพธ์ของ Factor โดยไม่มีการหมุนแกนปัจจัย โดยผลลัพธ์จะแสดงค่า communality , eigenvalues
- ✓ Scree plot แสดงกราฟค่า eigenvalues โดยเรียงลำดับจากมากไปน้อย โดยใช้ Factor ที่หมุนแกนปัจจัยแล้ว

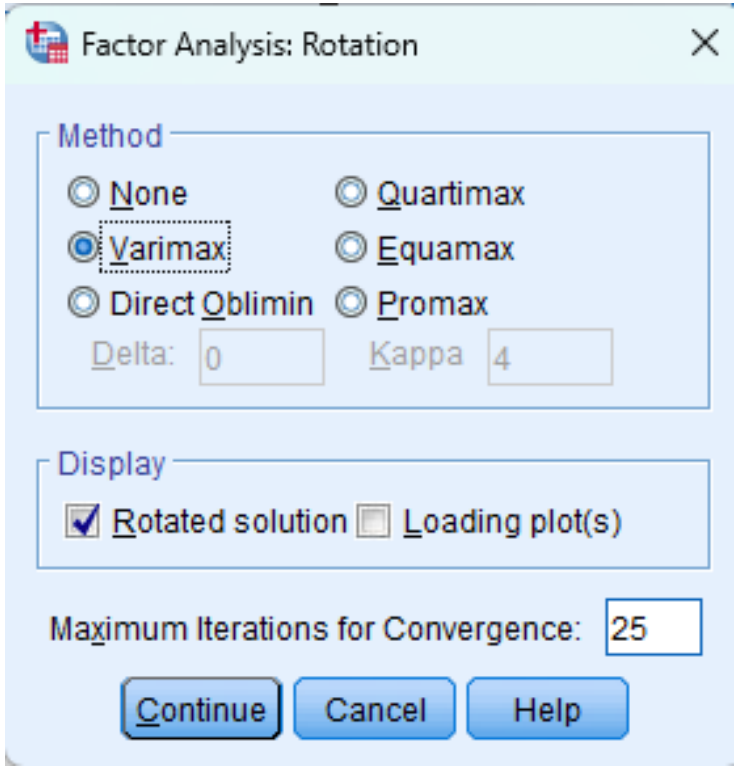
Extract

- Eigenvalues over: ระบุค่า eigenvalues = 1
- Number of factors: ใส่จำนวน Factor ที่ต้องการ

Maximum Iterations for Convergence

- กำหนดจำนวนรอบสูงสุดของการสกัดปัจจัยโดยโปรแกรม SPSS กำหนดเป็น 25 รอบ หรือเปลี่ยนมากกว่านั้นก็ได้หากข้อมูลขนาดใหญ่

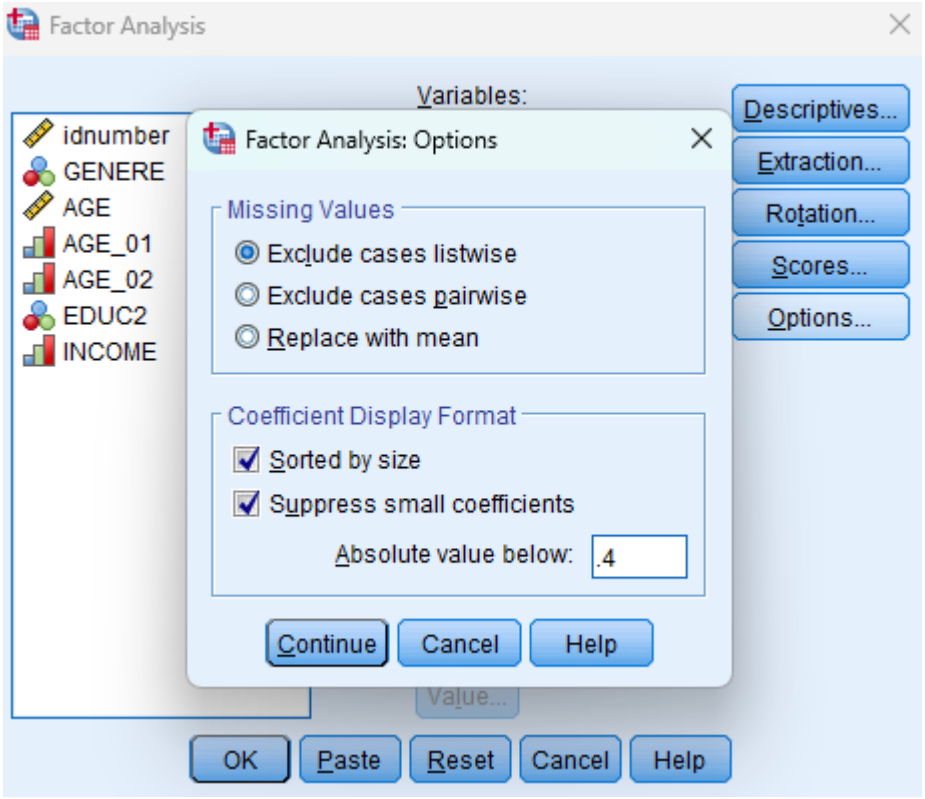
Factor analysis: Rotation



1. **Orthogonal Rotation เหมาะสำหรับ EFA หมุนแกนให้ factor ตั้งฉากกัน ทำให้ factors เป็นอิสระต่อกัน**
   1) Varimax เป็นเทคนิคที่ทำให้มีจำนวนตัวแปรที่น้อยที่สุด มีค่า Factor loading มากในแต่ละปัจจัย จึงเป็นวิธีที่นิยมใช้มากที่สุด
   2) Quartimax เป็นวิธีที่หมุนแกนปัจจัย โดยจะพยายามทำให้มีจำนวนปัจจัยน้อยที่สุด ในการอธิบายตัวแปรแต่ละตัว
   3) Equamax เป็นเทคนิคที่ใช้เกณฑ์ทั้งของ Varimax และ Quartimax
2. **Oblique Rotation เหมาะสำหรับ CFA หมุนแกนเป็นมุมแหลม ส่งผลให้ factors สัมพันธ์กัน**

   อนุญาตให้ factor ที่กำหนดไม่เป็นอิสระกัน
   1) Direct Oblimin
   2) Promax

Factor analysis: Options...



Missing

- Exclude case listwise จะวิเคราะห์เฉพาะ case ที่มีค่าของทุกตัวแปร
- Exclude case pairwise จะไม่รวม case ที่มี missing ของตัวแปรคู่ใดคู่หนึ่ง
- Replace with mean แทนค่า missing value ด้วยค่าเฉลี่ยของตัวแปรนั้น ๆ และใช้ทุก case ในการวิเคราะห์ปัจจัย

Coefficient Display Format แสดงค่าสัมประสิทธิ์

- Sorted by size จะแสดงค่า Factor loading เรียงตามลำดับ โดยตัวแปรที่มีค่า Factor loadingสูง ๆ ในปัจจัยเดียวกัน จะอยู่ด้วยกัน
- Suppress small coefficients
  - Absolute value below: ..... ระบุค่าที่ต้องการจะไม่แสดงค่าสัมประสิทธิ์สหสัมพันธ์ หรือ Factor loading ที่มีค่าน้อยกว่าที่ระบุ โดยค่าที่จะระบุมีค่า 0 ถึง 1 แนะนำ .3 ขึ้นไปเป็นอย่างน้อย

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 outlines the thresholds for sufficient/significant factor loadings.
- Generally, the smaller the sample size, the higher the required loading.
- 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.

The thresholds for sufficient/significant factor loadings

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

Output

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

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
HEALT	4.57	.671	735
BAEAUT	3.33	1.031	735
ANTIA	4.06	.878	735
LISTY	3.52	.920	735
TREND	3.16	.952	735
FSAFE	4.61	.661	735
POLIC	3.96	.967	735
TRADI	3.90	.858	735
AGRIC	3.71	.912	735
PRIND	3.56	.970	735
COUIS	3.60	.926	735

Correlation Matrix[a]

		HEALT	BAEAUT	ANTIA	LISTY	TREND	FSAFE	POLIC	TRADI	AGRIC	PRIND	COUIS
Correlation	HEALT	1.000	.260	.398	.015	.017	.408	.248	.101	.164	.195	.055
	BAEAUT	.260	1.000	.435	.225	.291	.155	.136	.212	.121	.139	.280
	ANTIA	.398	.435	1.000	.155	.185	.257	.200	.151	.183	.201	.154
	LISTY	.015	.225	.155	1.000	.411	.120	.119	.428	.256	.205	.429
	TREND	.017	.291	.185	.411	1.000	.123	.237	.346	.276	.211	.323
	FSAFE	.408	.155	.257	.120	.123	1.000	.378	.219	.291	.318	.139
	POLIC	.248	.136	.200	.119	.237	.378	1.000	.303	.421	.412	.087
	TRADI	.101	.212	.151	.428	.346	.219	.303	1.000	.408	.366	.468
	AGRIC	.164	.121	.183	.256	.276	.291	.421	.408	1.000	.772	.269
	PRIND	.195	.139	.201	.205	.211	.318	.412	.366	.772	1.000	.255
	COUIS	.055	.280	.154	.429	.323	.139	.087	.468	.269	.255	1.000

a. Determinant = .043

ค่า determinant > 0 แสดว่า การวิเคราะห์องค์ประกอบไม่มีปัญหา

Output: KMO and Barlett's Test

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.778
Bartlett's Test of Sphericity	Approx. Chi-Square	2293.830
	df	55
	Sig.	.000

Note:

- **Kaiser-Meyer-Olkin value**
  - มีค่าระหว่าง 0-1 ยิ่งเข้าใกล้ 1 ยิ่งดี
  - ค่าที่แนะนำคือ >.6
- **Bartlett's Test of Sphericity**
  - Bartlett's test should be less (p-value <.05).
  - This tests the null hypothesis that the correlation matrix is an identity matrix.
  - You want to reject this null hypothesis (Sig. < .05).

KMO statistics

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

Output: Communalities

Communalities แสดงค่าความร่วมกัน

- ค่า initial เป็นค่าเริ่มต้น/ค่าเดิมก่อนถูกแบ่งใน factor
- Extraction เป็นค่าที่อธิบายว่าตัวแปรนั้นสามารถอธิบายตัวแปรเดิมได้มากเพียงใด
  - ค่า Extraction ยิ่งมากยิ่งดี
  - ค่า Extraction น้อย สามารถเป็นตัวชี้วัดได้ว่า ตัวแปรนั้นควรปรับออกจากการวิเคราะห์ (ค่าระหว่าง 0.0-0.4)
- เช่น ตัวแปร Health หลังจากสกัดปัจจัยแล้ว ความแปรปรวนของตัวแปรถูกอธิบายโดยปัจจัยได้ 64.2%

Communalities

	Initial	Extraction
HEALT	1.000	.642
BAEAUT	1.000	.617
ANTIA	1.000	.625
LISTY	1.000	.586
TREND	1.000	.475
FSAFE	1.000	.492
POLIC	1.000	.510
TRADI	1.000	.566
AGRIC	1.000	.761
PRIND	1.000	.737
COUIS	1.000	.556

Extraction Method: Principal Component Analysis.

Component Matrix[a]

	Component		
	1	2	3
AGRIC	.718		-.493
PRIND	.700		-.480
TRADI	.669		
POLIC	.577		
COUIS	.558	-.467	
TREND	.546		
LISTY	.539	-.520	
FSAFE	.528	.461	
HEALT	.403	.626	
BAEAUT	.478		.623
ANTIA	.487		.521

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

ควรเลือกจำนวน factor เท่าไร

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	3.594	32.674	32.674	3.594	32.674	32.674	2.364	21.487	21.487
2	1.586	14.416	47.090	1.586	14.416	47.090	2.352	21.386	42.873
3	1.387	12.610	59.701	1.387	12.610	59.701	1.851	16.828	59.701
4	.803	7.301	67.001						
5	.778	7.072	74.073						
6	.602	5.475	79.548						
7	.561	5.104	84.652						
8	.515	4.679	89.331						
9	.494	4.494	93.824						
10	.458	4.161	97.985						
11	.222	2.015	100.000						

Extraction Method: Principal Component Analysis.

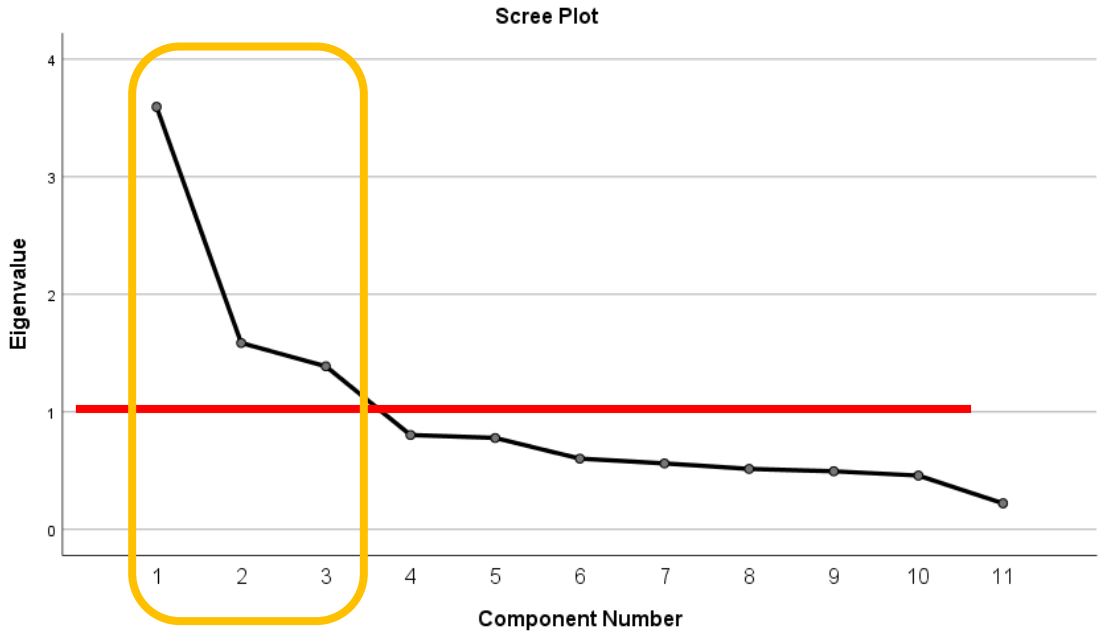
$$Initial\ Eigenvalue = \frac{total\ variance}{total\ no.\ of\ component}$$

Factor1 accounts for 32.67% of the total variance (3.594/11)

Factor2 accounts for 14.41% of the total variance (1.586/11)

Factor3 accounts for 12.61% of the total variance (1.383/11)

- Total หมายถึงค่า Eigenvalues ซึ่งคือความแปรปรวนทั้งหมดของตัวแปรเดิมที่อธิบายได้โดยปัจจัยนั้น ๆ เช่นปัจจัย 1 มีค่า Eigenvalues เท่ากับ 3.594 แสดงว่าปัจจัย 1 สามารถนำมาแทนตัวแปรเดิมได้ 3.594 ตัว (ดังนั้นจึงควรพิจารณาปัจจัยที่ Eigenvalues มากกว่า 1)
- 3 factors แรก สามารถอธิบายได้มากถึง 59.7% ของความแปรปรวนทั้งหมด
  - หมายความว่า ตัวแปรใหม่ที่ได้สามารถครอบคลุมข้อมูลเดิมของตัวแปรได้ 59.7%



Output: Component and Rotated Component Matrix

Component Matrix[a]

	Component		
	1	2	3
AGRIC	.718		-.493
PRIND	.700		-.480
TRADI	.669		
POLIC	.577		
COUIS	.558	-.467	
TREND	.546		
LISTY	.539	-.520	
FSAFE	.528	.461	
HEALT	.403	.626	
BAEAUT	.478		.623
ANTIA	.487		.521

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

Rotated Component Matrix[a]

	Component		
	1	2	3
PRIND	.828		
AGRIC	.822		
POLIC	.676		
FSAFE	.521		.469
LISTY		.761	
COUIS		.735	
TREND		.668	
TRADI		.643	
ANTIA			.767
HEALT			.740
BAEAUT		.429	.651

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

- ค่าที่แสดงเป็นค่า Factor Loading หรือค่าที่แสดงความสัมพันธ์ระหว่างตัวแปรแต่ละตัวกับ Factor
- เช่น ตัวแปร FSAFE มีค่า Factor Loading ของปัจจัยแรกเท่ากับ .521 มากกว่าค่า Factor Loading ของปัจจัยที่ 3 ที่ .469
- Factor loading พิจารณาเปรียบเทียบเฉพาะค่า absolute ไม่พิจารณาเครื่องหมาย

Component Transformation Matrix

Component	1	2	3
1	.658	.623	.424
2	.329	-.744	.582
3	-.678	.243	.694

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

การตั้งชื่อตัวแปร

Rotated Component Matrix[a]

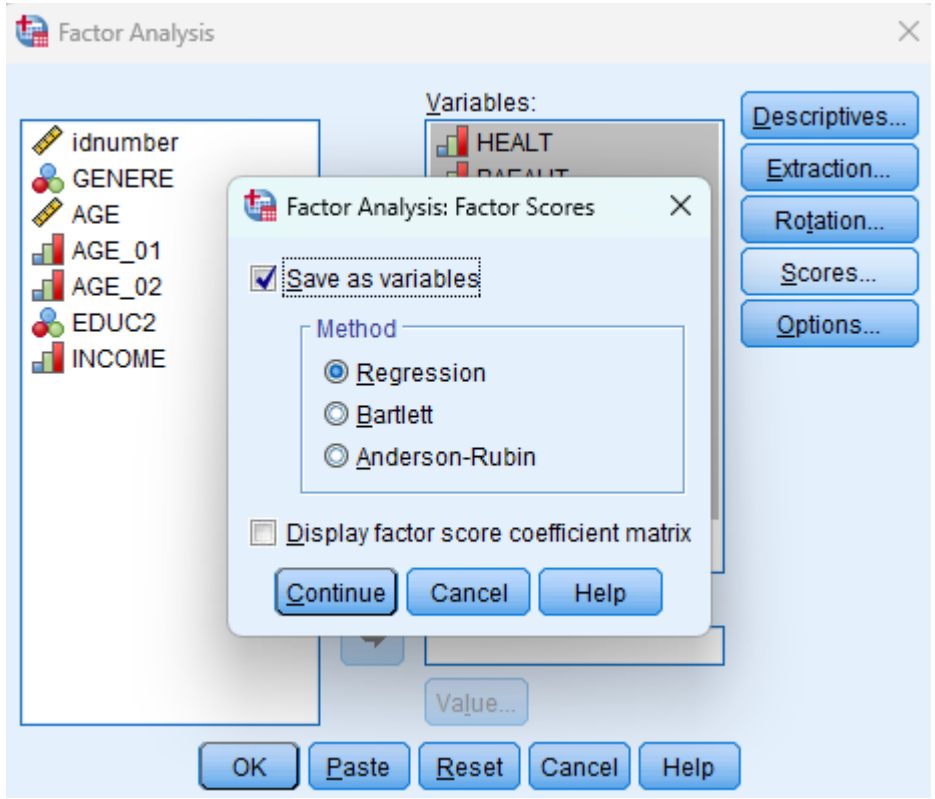
	Component		
	1	2	3
PRIND	.828		
AGRIC	.822		
POLIC	.676		
FSAFE	.521		.469
LISTY		.761	
COUIS		.735	
TREND		.668	
TRADI		.643	
ANTIA			.767
HEALT			.740
BAEAUT		.429	.651

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

Information about the food processing industry and innovations in terms of products and processes	PRIND
Information about the production techniques used in the primary sector	AGRIC
Information about food regulations, affecting consumer choices and the food industry	POLIC
Information about food safety issues caused by bacteria and other substances	FSAFE
Information about lifestyle, food tourism and eating out	LISTY
Information about Italian and international cuisine, food culture and good living	COUIS
information about trends, consumption evolution, food fads, and underscoring ethnicity, cultural, social diversity of Italian polulation	TREND
Information about tradition, regional typical products and quality foods that are disappearing from the litalian market	TRADI
Information about food containing anti ageing properties	ANTIA
Information about health risks caused by obesity, anorexia nervosa bulimia and other illesses liked to food	HEALT
Information about foods related to health and beauty	BEAUT

Factor analysis: Scores...



Save as variables

- เมื่อเลือกทางเลือกนี้จะเป็นการ save Factor score ในรูปของตัวแปร (1 factor = 1 new variable)
- Factor score มีวิธีการคำนวณให้เลือก 3 วิธี
  - Regression โดยวิธีนี้ให้ค่าแปรปรวนเท่ากับ (สัมประสิทธิ์สหสัมพันธ์ระหว่างค่า Factor score ที่ประมาณได้กับค่า Factor score จริง)
  - Bartlett
  - Anderson-Rubin

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

การวิเคราะห์จัดกลุ่ม (Cluster analysis)

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

BA603

ความหมายของการวิเคราะห์จัดกลุ่ม

- การจัด case: คน สัตว์ สิ่งของ หรือองค์กร ฯลฯ เป็นการจัดตัวแปรออกเป็นกลุ่มย่อย ตั้งแต่ 2 กลุ่มขึ้นไป
  - กลุ่ม (cluster) เดียวกันจะมี case คล้าย ๆ กัน
  - case ที่ต่างกันจะอยู่คนละกลุ่มกัน
- นิยมใช้จัดกลุ่มและสร้าง profile of reponses

ข้อตกลงเบื้องต้นสำหรับการวิเคราะห์จัดกลุ่ม

- ไม่ทราบจำนวนมาก่อนว่ามีกี่กลุ่ม
- ไม่ทราบมาก่อนว่าใครอยู่กลุ่มใด
- คนหนึ่งคนอยู่ได้เพียงกลุ่มเดียว
- ใช้หลายตัวแปรเป็นปัจจัยในการแบ่งกลุ่ม

Methods

เป็นขั้นตอน

Hierarchical procedure

Hierarchical cluster

- is the most common method. It generates a series of models with cluster solutions from 1 (all cases in one cluster) to n (each case is an individual cluster).
- Hierarchical cluster also works with variables as opposed to cases; it can cluster variables together in a manner somewhat similar to factor analysis. In addition, hierarchical cluster analysis can handle nominal, ordinal, and scale data; however, it is not recommended to mix different levels of measurement.

เป็นขั้นตอน

Non - hierarchical procedure

K-means cluster

- is a method to quickly cluster large data set. The researcher define the number of clusters in advance. This is useful to test different models with a different assumed number of clusters

Two-step cluster

- analysis identifies groupings by running pre-clustering first and then by running hierarchical methods. Because it uses a quick cluster algorithm upfront, it can handle large data sets that would take a long time to compute with hierarchical cluster methods. In this respect, it is a combination of the previous two approaches. Two-step clustering can handle scale and ordinal data in the same model, and it automatically selects the number of clusters.

Hierarchical vs Non hierarchical methods

Hierarchical clustering

- No decision about the number of clusters
- Problems when data contain a high level of error
- Can be very slow
- Initial decision are more influential (one-step only)

Non hierarchical clustering

- Faster, more reliable
- Need to specify the number of clusters (arbitrary)
- Need to set the initial seeds (arbitrary)

Suggested approach

1. First perform a hierarchical method to define the number of clusters
2. Then use the k-means procedure to form the clusters

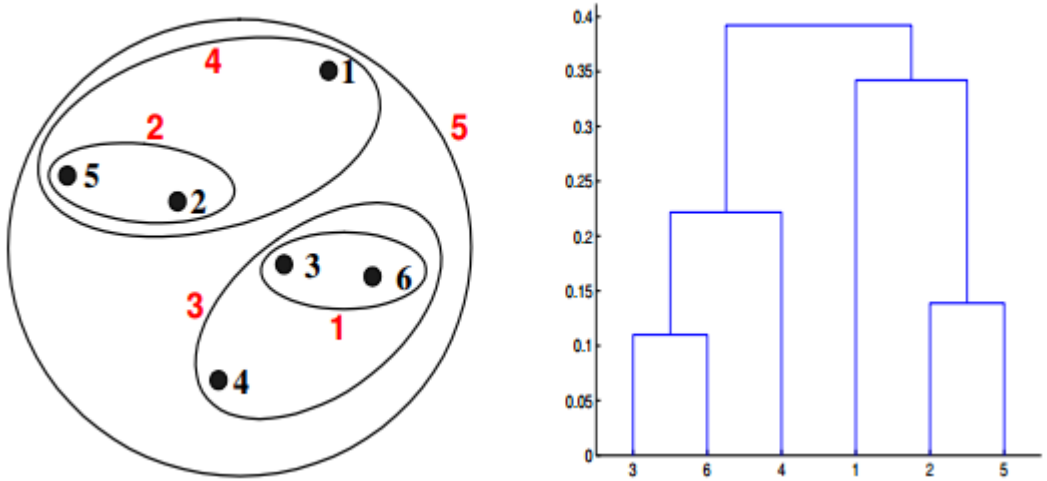
ขั้นตอนในการวิเคราะห์การจัดกลุ่ม

1. เลือกวิธีวัดระยะห่าง (Distance measure)
2. เลือกวิธีการจัดกลุ่ม (Clustering algorithm)
3. ระบุจำนวนกลุ่ม (Determine the number of clusters)
4. ตรวจสอบความเหมาะสมของผลการวิเคราะห์ (Validate the analysis)

Cluster analysis: basic steps

1. Apply Ward's methods on the principal components score
2. Check the agglomeration schedule
3. Decide the number of clusters
4. Apply the k-means method

Hierarchical cluster analysis: HCA

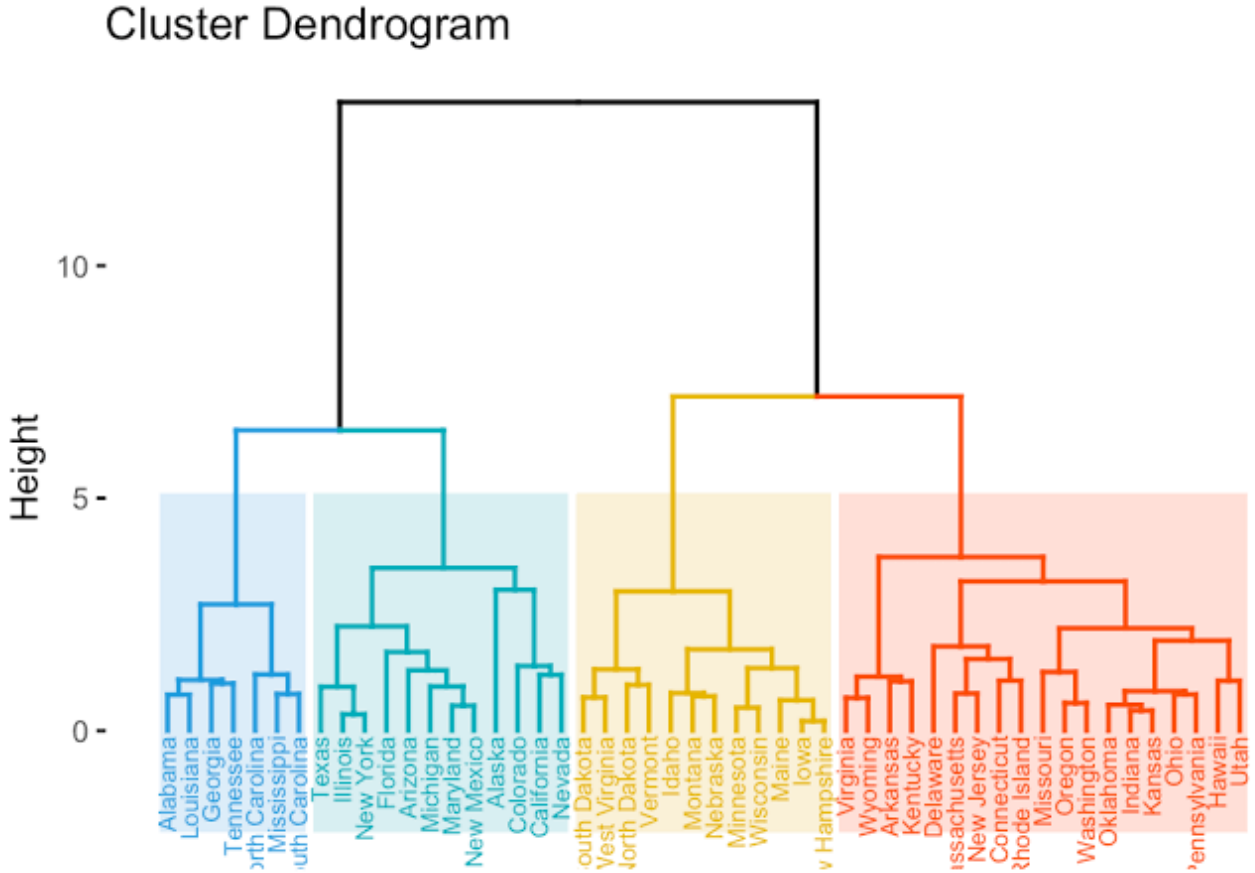


การวิเคราะห์จัดกลุ่มตามลำดับชั้น เป็นการจัดกลุ่มที่นิยมใช้แบ่งกลุ่ม Case หรือกลุ่มตัวแปร โดยมีเงื่อนไข ดังนี้

1. เหมาะกับข้อมูลขนาดเล็ก (จำนวนเคส < 200 เคส/ตัวแปร)
   - ตัวอย่าง => แบ่งกลุ่ม (Classify cases)
   - ตัวแปร => ทดสอบความสัมพันธ์ระหว่างตัวแปร
2. ไม่จำเป็นต้องทราบจำนวนกลุ่มมาก่อน
3. ไม่จำเป็นต้องทราบว่าตัวแปรใดหรือเคสใดอยู่กลุ่มใดก่อน
4. ชนิดตัวแปรที่เหมาะสมคือ nominal, ordinal, and scale data และไม่ควรผสมชนิดของตัวแปร

HCA: เกณฑ์ในการจัดกลุ่ม

แยกแต่ละตัวอย่างไปยังกลุ่ม ด้วยระยะทาง (Distance) เริ่มต้นโดยแบ่งแยก 2 กลุ่มที่เหมือนกันมากที่สุดไปเรื่อย ๆ จนครบทุกตัวอย่าง

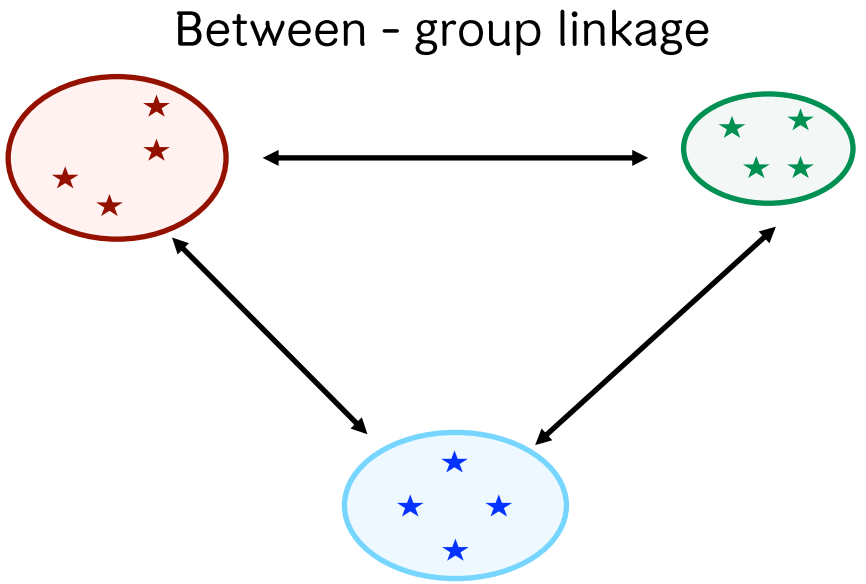


ขั้นตอนของเทคนิค Hierarchical cluster analysis สำหรับการแบ่งกลุ่มเคส

1. เลือกตัวแปรหรือปัจจัยที่คาดว่ามีอิทธิพลที่ทำให้เคสต่างกัน ตัวแปรจะทำให้สามารถแบ่งกลุ่มเคสได้ชัดเจน
2. เลือกวิธีการวัดระยะห่างระหว่างเคสแต่ละคู่ หรือเลือกวิธีการคำนวณเพื่อวัดค่า ความคล้ายของเคสแต่ละคู่
3. เลือกเกณฑ์ในการรวมกลุ่มหรือรวม Cluster

HCA: เกณฑ์ในการจัดกลุ่ม

1. **Between - group linkage** (or average linkage between group)
2. **Within-group linkage** (or average linkage within groups method) - วิธีนี้จะรวม cluster เข้าด้วยกันถ้าระยะห่างเฉลี่ยระหว่างทุกเคสใน cluster นั้นๆ มีค่าน้อยที่สุด
3. **Centroid clustering** - รวม 2 cluster เข้าด้วยกัน โดยพิจารณาจากระยะห่างของจุดกลาวของ cluster 2 cluster
4. **Ward's method** - พิจารณาค่า sum of the squared within-cluster distance โดยจะรวม cluster ที่ทำให้ค่า sum of the squared within-cluster distance เพิ่มขึ้นน้อยที่สุด โดยค่า square within-cluster distance คือค่า square Euclidean distance ของแต่ละเคสกับ cluster mean



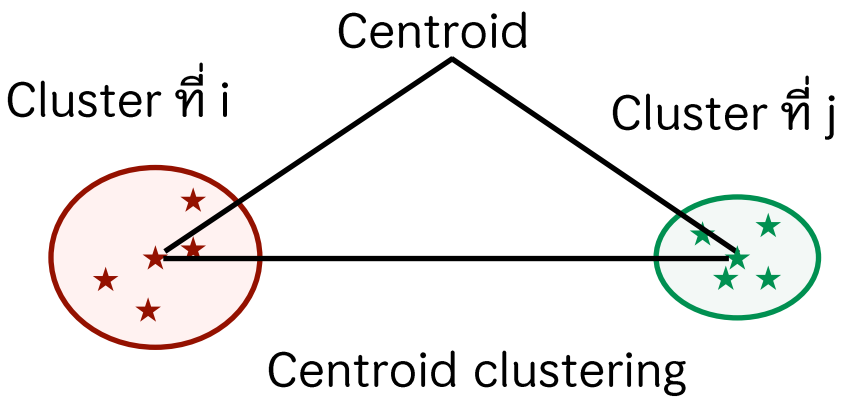
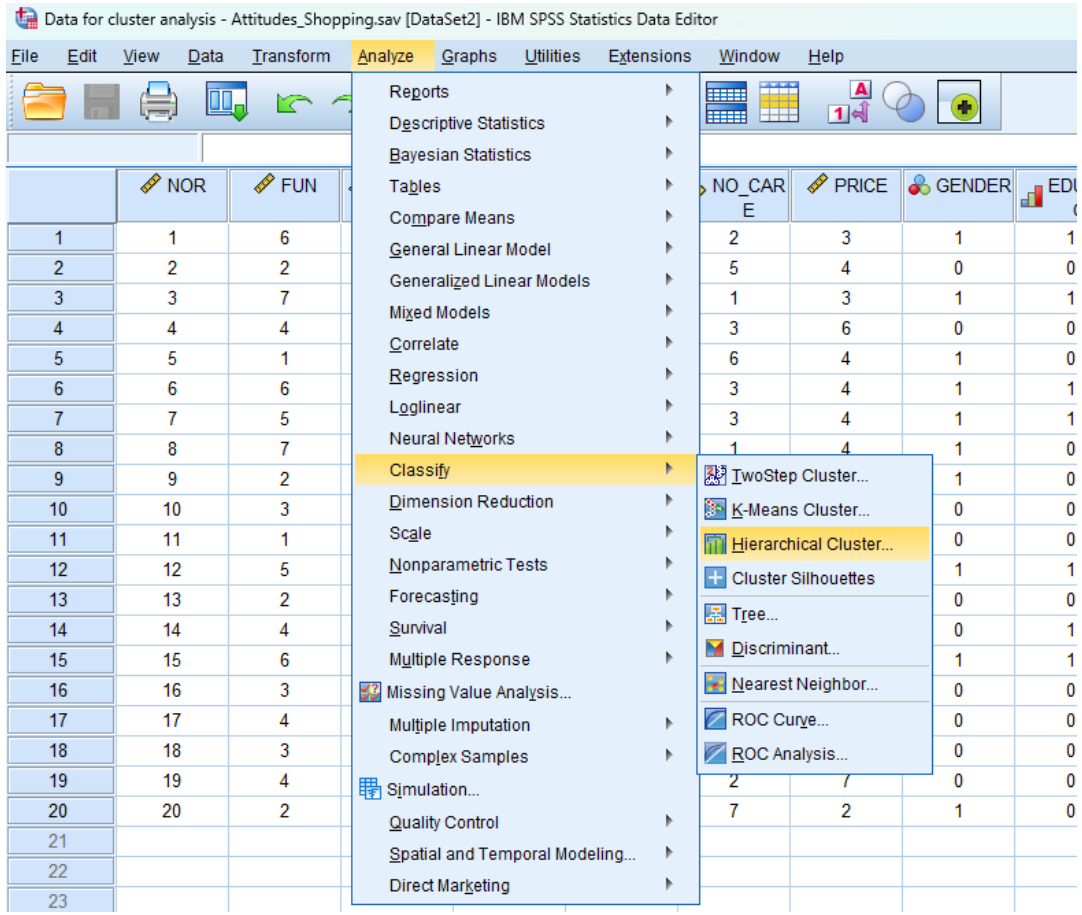
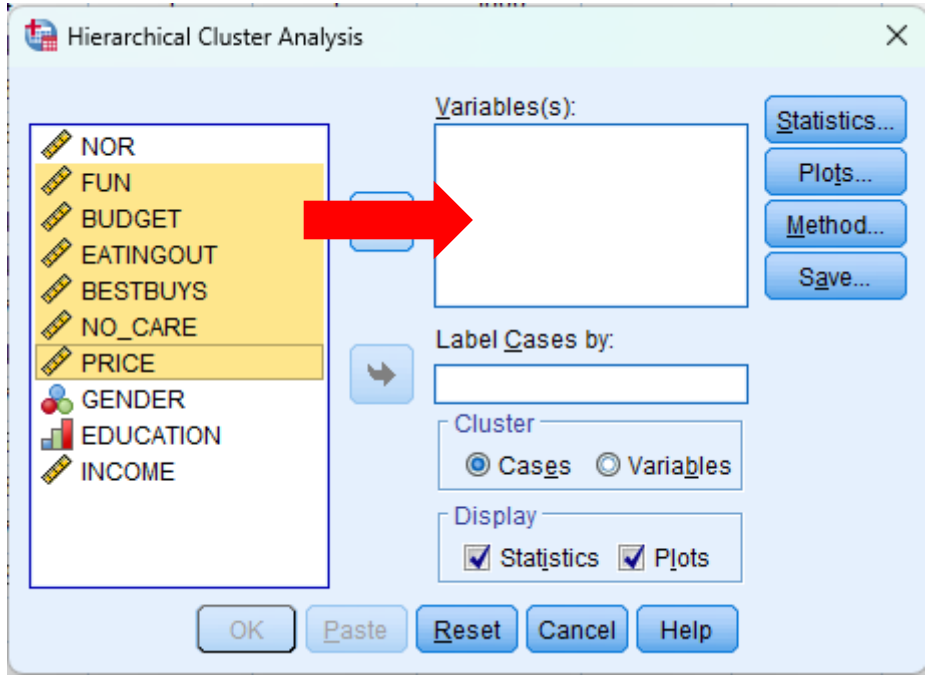


Table 1: Items used to measure consumers attitudes towards shopping.

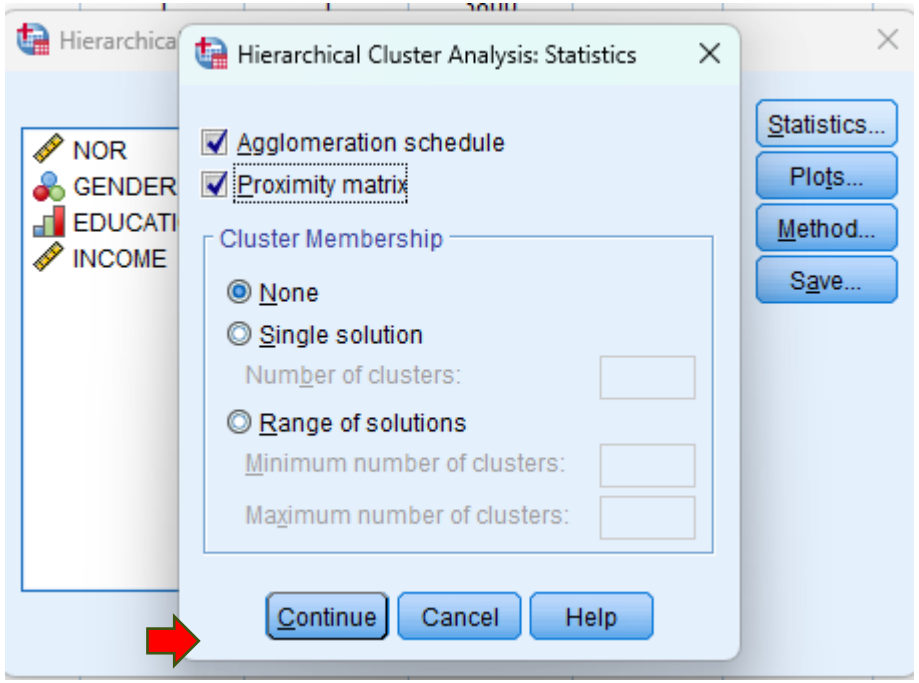
Variables name	Attitude items	Type of data
1. NOR	Number of respondent	Scale
2. FUN	Shopping is fun	Scale
3. BUDGET	Shopping is bud for your budget	Scale
4. EATINGOUT	I combine shopping with eating out	Scale
5. BESTBUYS	I try to get the best buys when shopping	Scale
6. NO_CARE	I don’t care about shopping	Scale
7. PRICE	You can save a lot of money by comparing prices	Scale
8. GENDER		Nominal
9. EDUCATION		Ordinal
10. INCOME		Scale

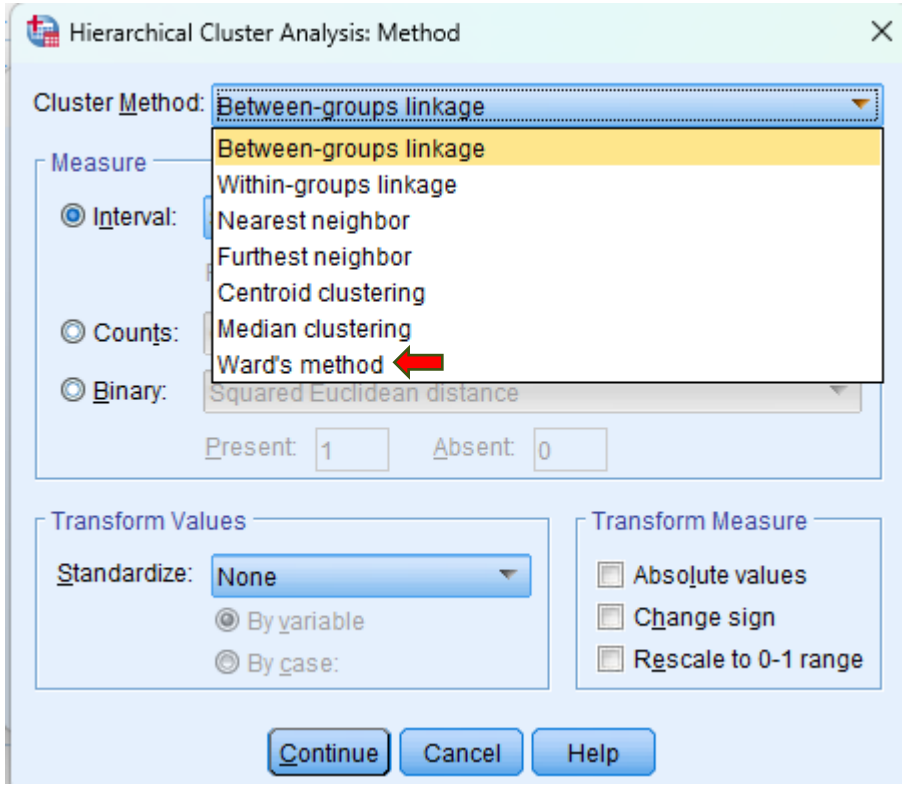
Analyze > Classify > Hierarchical cluster…





HCA: Statistics





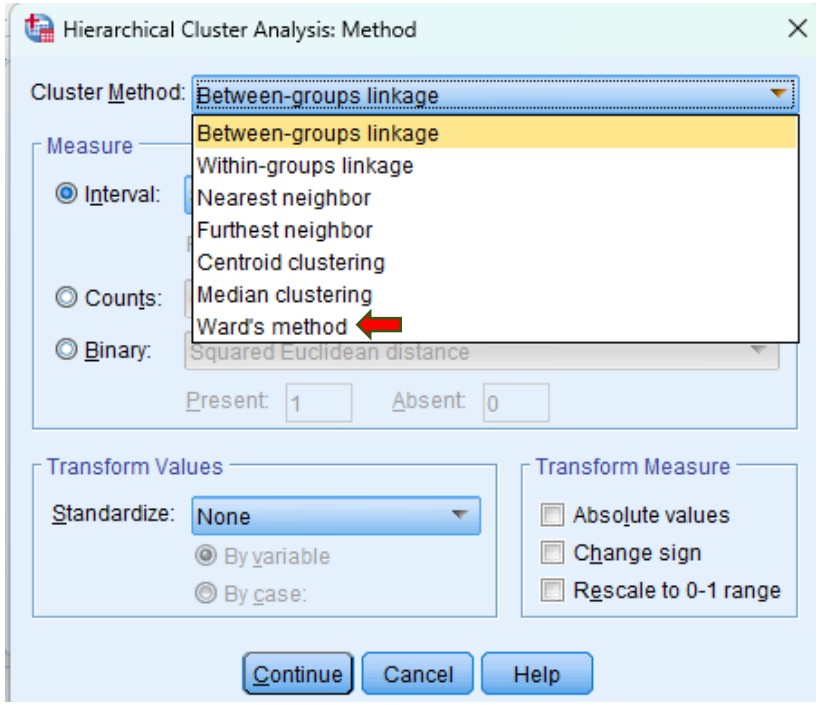
Agglomerative clustering method

Linkage methods

- Single linkage (minimum distance)
- Complete linkage (maximum distance)
- Average linkage

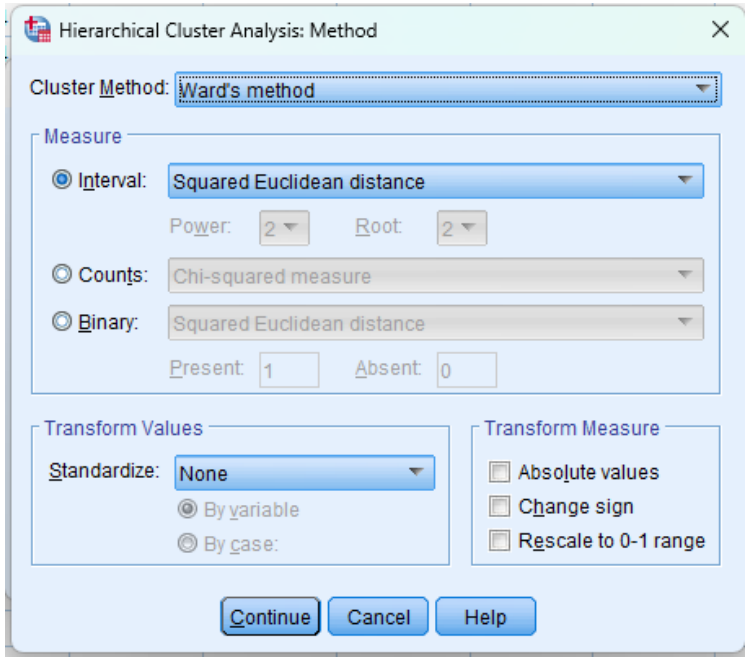
Centroid method

- The distance between two clusters is defined as the difference between the centroids (cluster averages)



Ward's method

1. Compute sum of squared distances within clusters
2. Aggregate clusters with the minimum increase in the overall sum of squares



Proximity Matrix: Square Euclidean Distance

Proximity Matrix

Squared Euclidean Distance

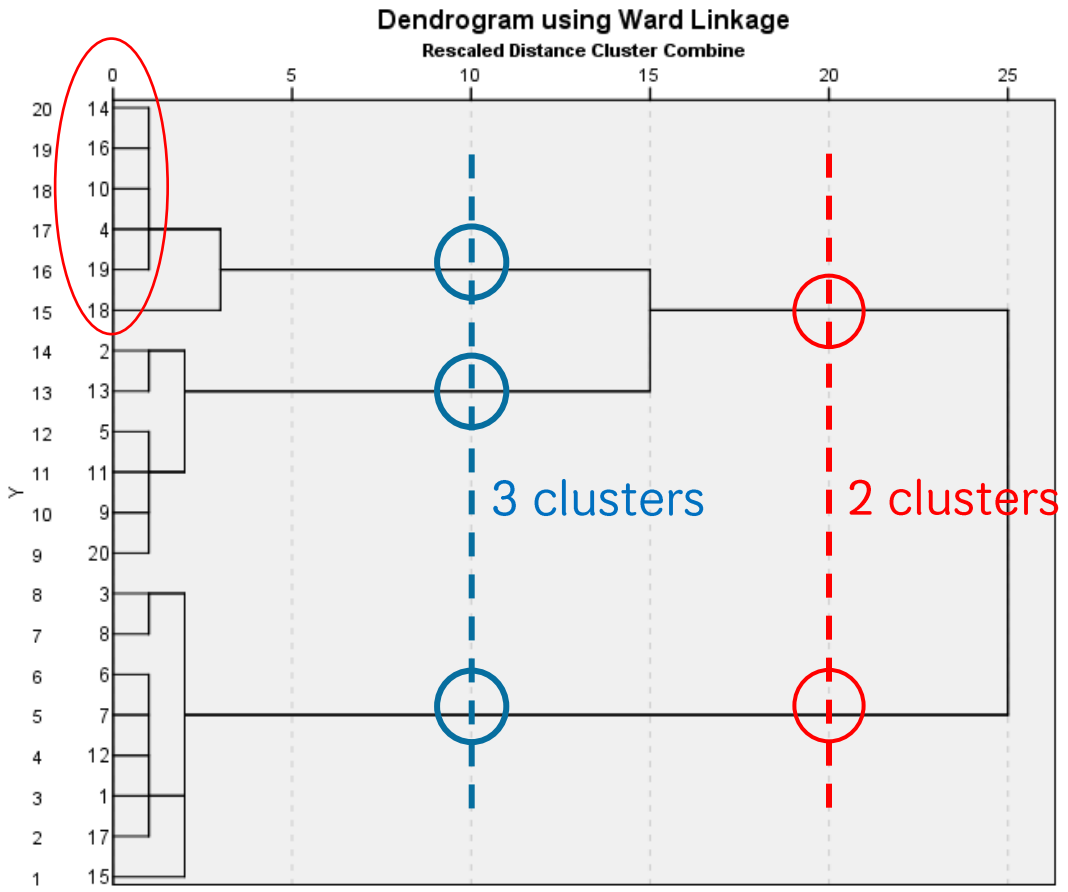
Case	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	.000	64.000	8.000	31.000	69.000	3.000	5.000	5.000	48.000	48.000	60.000	7.000	65.000	46.000	13.000	48.000	9.000	56.000	56.000	69.000
2	64.000	.000	68.000	31.000	7.000	47.000	39.000	77.000	8.000	18.000	4.000	35.000	3.000	36.000	49.000	28.000	55.000	24.000	44.000	9.000
3	8.000	68.000	.000	43.000	83.000	11.000	11.000	3.000	64.000	56.000	70.000	11.000	61.000	58.000	19.000	58.000	23.000	70.000	60.000	79.000
4	31.000	31.000	43.000	.000	44.000	20.000	22.000	36.000	31.000	5.000	39.000	12.000	34.000	3.000	22.000	5.000	24.000	17.000	7.000	50.000
5	69.000	7.000	83.000	44.000	.000	52.000	42.000	90.000	5.000	33.000	3.000	46.000	16.000	51.000	58.000	41.000	52.000	41.000	69.000	10.000
6	3.000	47.000	11.000	20.000	52.000	.000	2.000	8.000	35.000	33.000	47.000	4.000	50.000	31.000	10.000	33.000	8.000	45.000	43.000	54.000
7	5.000	39.000	11.000	22.000	42.000	2.000	.000	10.000	29.000	31.000	37.000	4.000	40.000	33.000	14.000	31.000	6.000	47.000	45.000	46.000
8	5.000	77.000	3.000	36.000	90.000	8.000	10.000	.000	69.000	53.000	79.000	10.000	72.000	49.000	18.000	51.000	16.000	71.000	53.000	90.000
9	48.000	8.000	64.000	31.000	5.000	35.000	29.000	69.000	.000	24.000	4.000	31.000	17.000	38.000	45.000	32.000	41.000	24.000	56.000	5.000
10	48.000	18.000	56.000	5.000	33.000	33.000	31.000	53.000	24.000	.000	28.000	21.000	19.000	4.000	39.000	2.000	39.000	14.000	8.000	35.000
11	60.000	4.000	70.000	39.000	3.000	47.000	37.000	79.000	4.000	28.000	.000	37.000	9.000	48.000	51.000	38.000	49.000	30.000	60.000	7.000
12	7.000	35.000	11.000	12.000	46.000	4.000	4.000	10.000	31.000	21.000	37.000	.000	34.000	23.000	8.000	23.000	10.000	31.000	27.000	48.000
13	65.000	3.000	61.000	34.000	16.000	50.000	40.000	72.000	17.000	19.000	9.000	34.000	.000	39.000	52.000	29.000	58.000	29.000	41.000	16.000
14	46.000	36.000	58.000	3.000	51.000	31.000	33.000	49.000	38.000	4.000	48.000	23.000	39.000	.000	39.000	2.000	37.000	22.000	6.000	55.000
15	13.000	49.000	19.000	22.000	58.000	10.000	14.000	18.000	45.000	39.000	51.000	8.000	52.000	39.000	.000	43.000	16.000	43.000	41.000	68.000
16	48.000	28.000	58.000	5.000	41.000	33.000	31.000	51.000	32.000	2.000	38.000	23.000	29.000	2.000	43.000	.000	35.000	24.000	8.000	47.000
17	9.000	55.000	23.000	24.000	52.000	8.000	6.000	16.000	41.000	39.000	49.000	10.000	58.000	37.000	16.000	35.000	.000	59.000	49.000	68.000
18	56.000	24.000	70.000	17.000	41.000	45.000	47.000	71.000	24.000	14.000	30.000	31.000	29.000	22.000	43.000	24.000	59.000	.000	24.000	31.000
19	56.000	44.000	60.000	7.000	69.000	43.000	45.000	53.000	56.000	8.000	60.000	27.000	41.000	6.000	41.000	8.000	49.000	24.000	.000	73.000
20	69.000	9.000	79.000	50.000	10.000	54.000	46.000	90.000	5.000	35.000	7.000	48.000	16.000	55.000	68.000	47.000	68.000	31.000	73.000	.000

This is a dissimilarity matrix

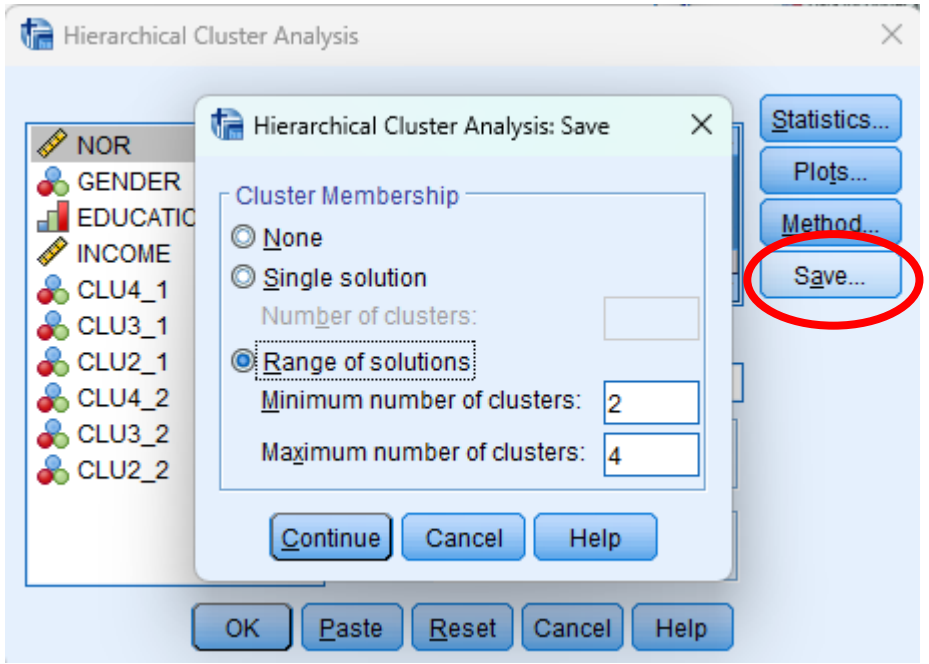
Ward Linkage

Agglomeration Schedule

Stage	Cluster Combined Cluster 1	Cluster Combined Cluster 2	Coefficients	Stage Cluster First Appears Cluster 1	Stage Cluster First Appears Cluster 2	Next Stage
1	14	16	1.000	0	0	6
2	6	7	2.000	0	0	7
3	2	13	3.500	0	0	15
4	5	11	5.000	0	0	11
5	3	8	6.500	0	0	16
6	10	14	8.167	0	1	9
7	6	12	10.500	2	0	10
8	9	20	13.000	0	0	11
9	4	10	15.583	0	6	12
10	1	6	18.500	0	7	13
11	5	9	23.000	4	8	15
12	4	19	27.750	9	0	17
13	1	17	33.100	10	0	14
14	1	15	41.333	13	0	16
15	2	5	51.833	3	11	18
16	1	3	64.500	14	5	19
17	4	18	79.667	12	0	18
18	2	4	172.667	15	17	19
19	1	2	328.600	16	18	0



Save membership



*Data for cluster analysis - Attitudes_Shopping.sav [DataSet2] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Extensions Window

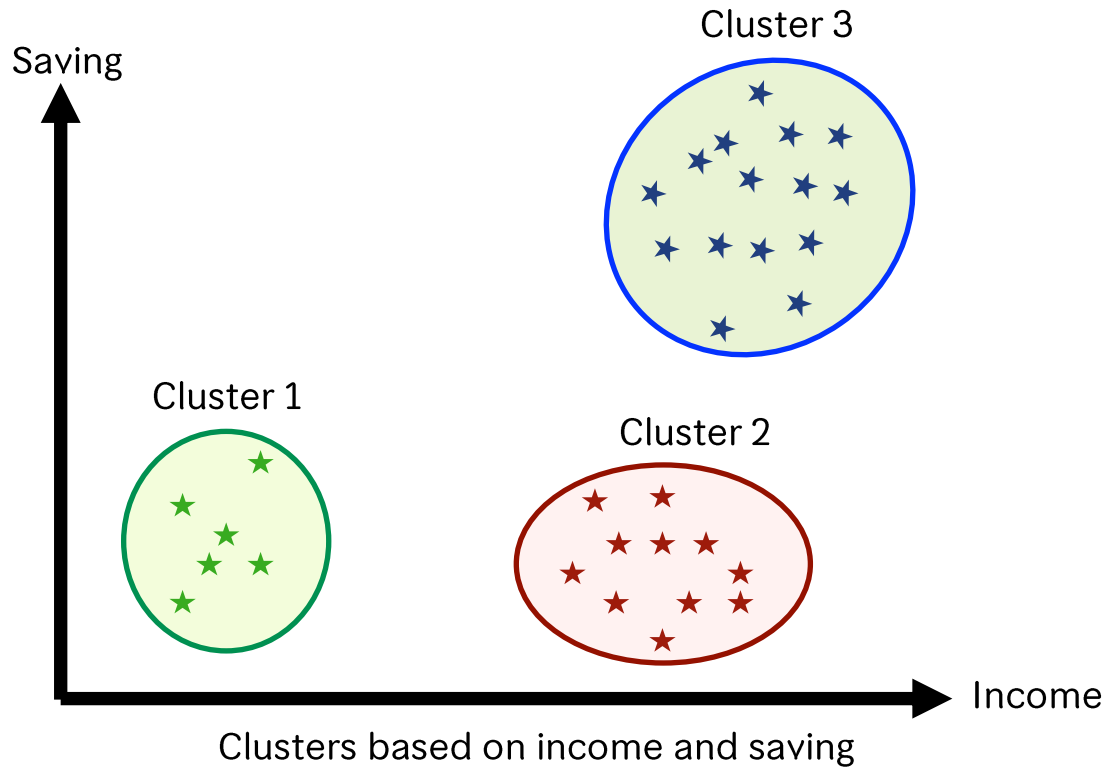
	Name	Type	Width	Decimals	Label	Values
1	NOR	Numeric	8	0	Number of resp...	None
2	FUN	Numeric	8	0	Shopping is fun	{1, Extremel...
3	BUDGET	Numeric	8	0	Shopping is bu...	{1, Extremel...
4	EATINGOUT	Numeric	8	0	I combine shop...	{1, Extremel...
5	BESTBUYS	Numeric	8	0	I try to get the ...	{1, Extremel...
6	NO_CARE	Numeric	8	0	I don't care abo...	{1, Extremel...
7	PRICE	Numeric	8	0	You can save a...	{1, Extremel...
8	GENDER	Numeric	8	0	Gender	None
9	EDUCATION	Numeric	8	0	Education	None
10	INCOME	Numeric	8	0	Net monthly inc...	None
11	CLU4_1	Numeric	8	0	Ward Method	None
12	CLU3_1	Numeric	8	0	Ward Method	None
13	CLU2_1	Numeric	8	0	Ward Method	None
14						

ตัวแปรที่จัดกลุ่มๆละ
2-4 กลุ่มจะปรากฏขึ้น

K-means cluster analysis: KCA

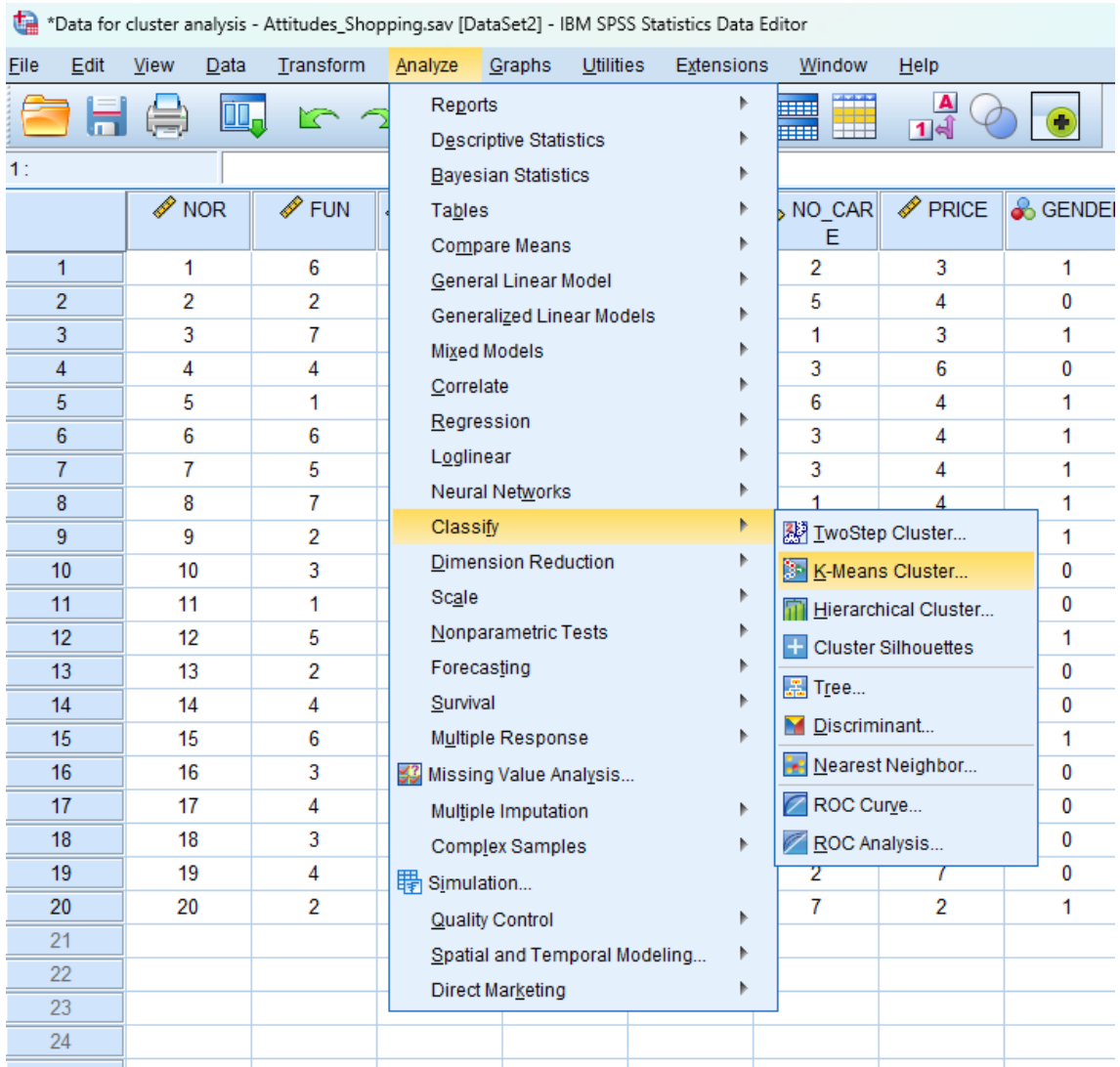
KCA is a method to quickly cluster large data sets. **The researcher define the number of clusters in advance**. This is useful to test different models with a different assumed number of clusters.

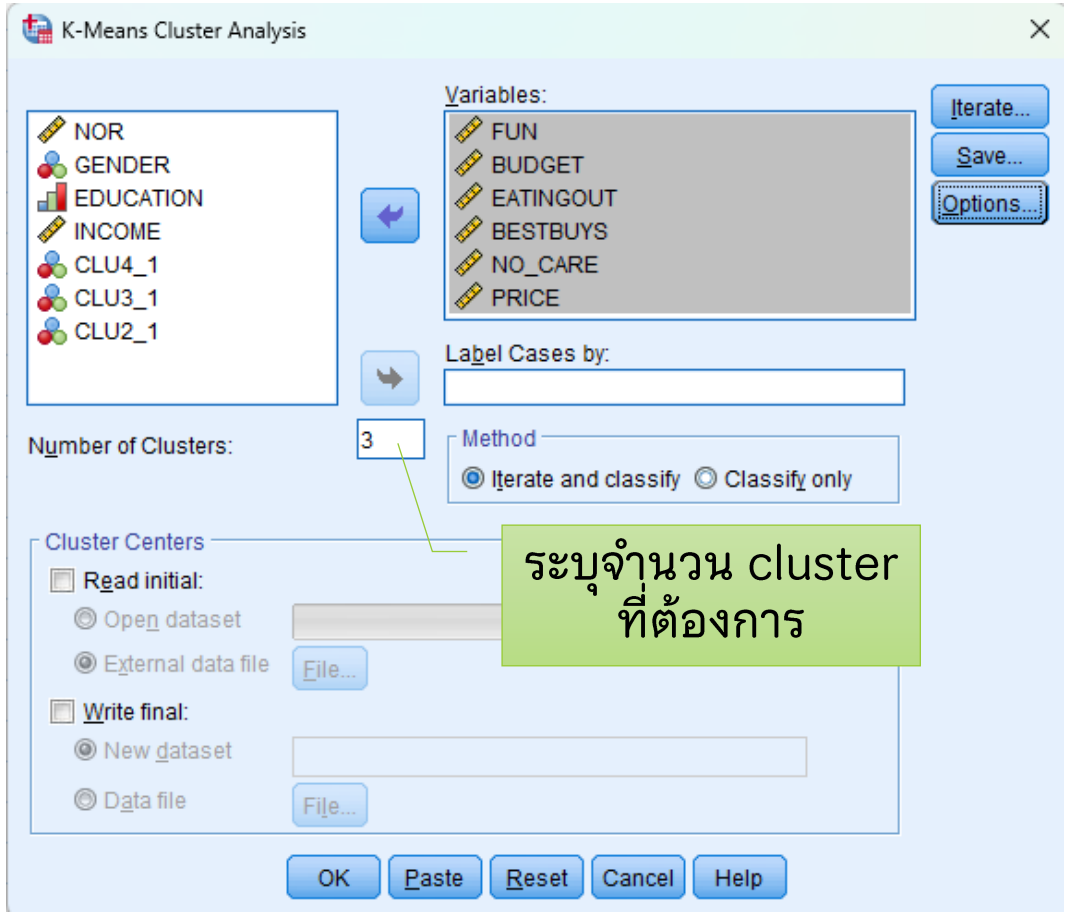
- กำหนดจำนวนกลุ่ม (K) ที่ต้องการก่อน แล้วจัดตัวอย่างเข้ากลุ่ม
- ใช้กับข้อมูลขนาดใหญ่ (> 200)
- เหมาะกับตัวแปรค่าต่อเนื่อง (Scale)
- ใช้ค่าเฉลี่ยเพื่อจัด Case เข้ากลุ่ม K
- กลุ่มต่างกันน้อย กลุ่มต่างกันมาก
- หาระยะห่างด้วยวิธี Euclidean distance



Note: **หากมีตัวแปรที่มีการแจกแจงไม่ปกติ ให้ทำการแปลงเป็นค่ามาตรฐานก่อน (Z-score)**

KCA: Analyze > Classify > K-means Cluster…





KCA: Output

Initial Cluster Centers

	Cluster		
	1	2	3
Shopping is fun	4	2	7
Shopping is bud for your budget	6	3	2
I combine shopping with eating out	3	2	6
I try to get the best buys when shopping	7	4	4
I don't care about shopping	2	7	1
You can save a lot of money by comparing prices	7	2	3

Distances between Final Cluster Centers

Cluster	1	2	3
1		5.568	5.698
2	5.568		6.928
3	5.698	6.928	

Cluster Membership

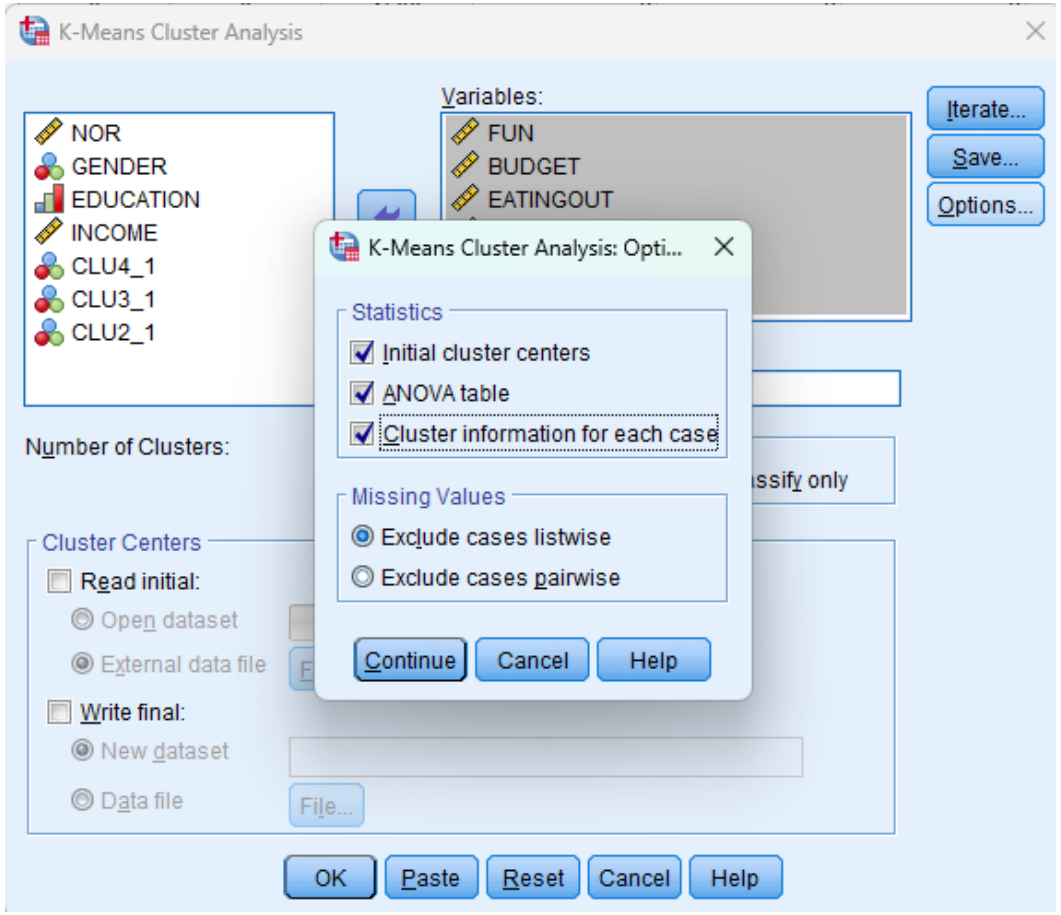
Case Number	Cluster	Distance
1	3	1.414
2	2	1.323
3	3	2.550
4	1	1.404
5	2	1.848
6	3	1.225
7	3	1.500
8	3	2.121
9	2	1.756
10	1	1.143
11	2	1.041
12	3	1.581
13	2	2.598
14	1	1.404
15	3	2.828
16	1	1.624
17	3	2.598
18	1	3.555
19	1	2.154
20	2	2.102

Number of Cases in each Cluster

Cluster	1	6.000
	2	6.000
	3	8.000
Valid		20.000
Missing		.000

แสดงจำนวนตัวอย่างของแต่ละกลุ่ม

KCA: Options



ANOVA table

- ใช้ชี้วัดว่าตัวแปรใดมีผลมากที่สุดต่อการกำหนดกลุ่มที่กำหนดไว้
- ตัวแปรที่มีค่า F-stat สูงสุด แสดงว่าเป็นตัวแปรที่ใช้แบ่งกลุ่มได้ดีที่สุด

ANOVA

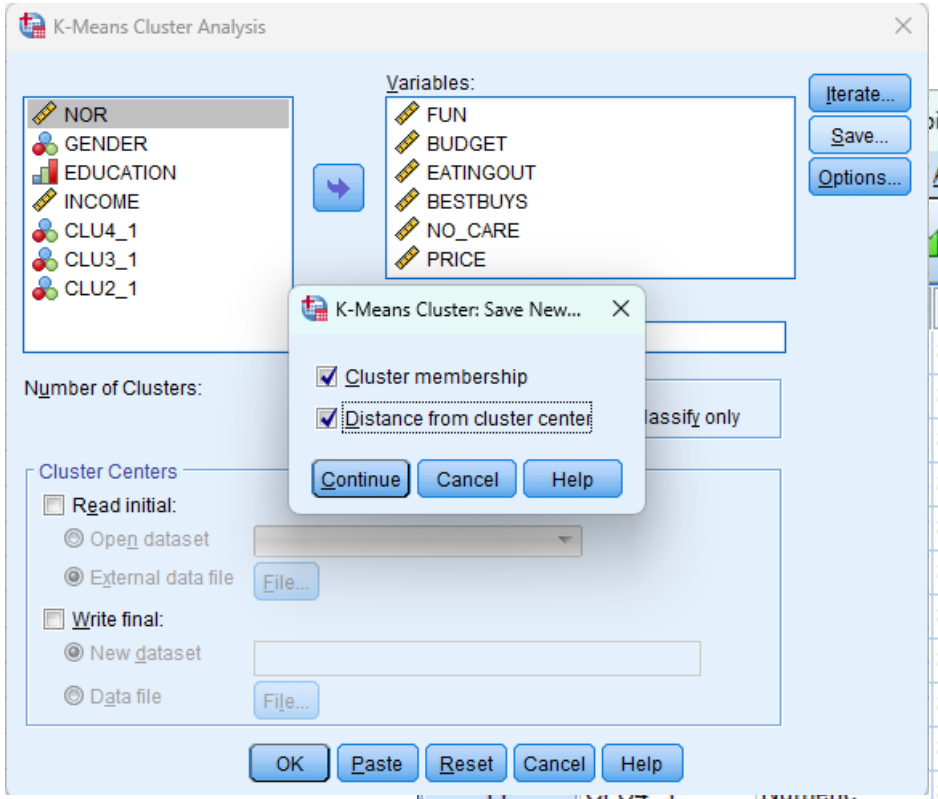
	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Shopping is fun	29.108	2	.608	17	47.888	.000
Shopping is bud for your budget	13.546	2	.630	17	21.505	.000
I combine shopping with eating out	31.392	2	.833	17	37.670	.000
I try to get the best buys when shopping	15.713	2	.728	17	21.585	.000
I don't care about shopping	22.537	2	.816	17	27.614	.000
You can save a lot of money by comparing prices	12.171	2	1.071	17	11.363	.001

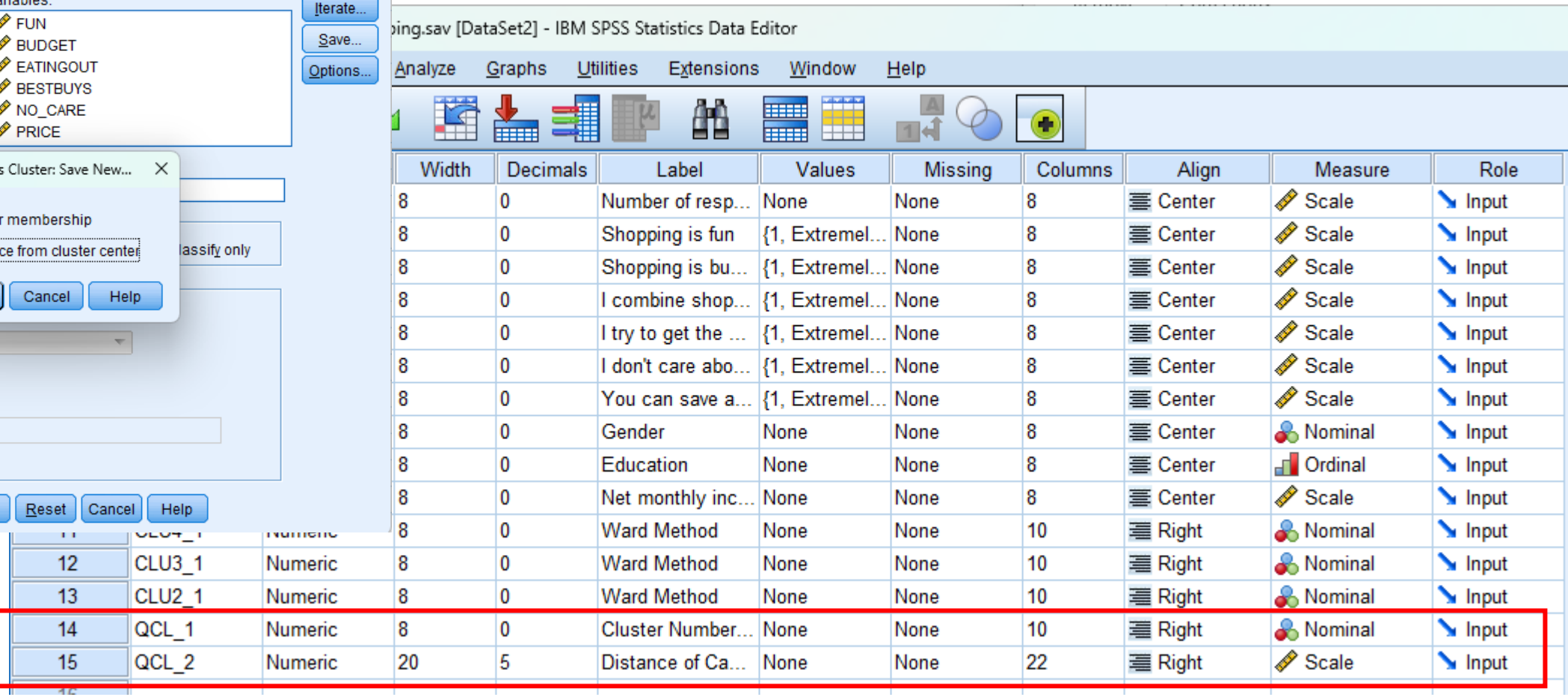
The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

KCA: Save

ได้ตัวแปร 2 ตัว

- QCL1 ตัวแปรที่กำหนดว่าใครอยู่กลุ่มใคร
- QCL2 ตัวแปรที่บอกระยะห่างตัวแต่ละเคสจากคลัสเตอร์



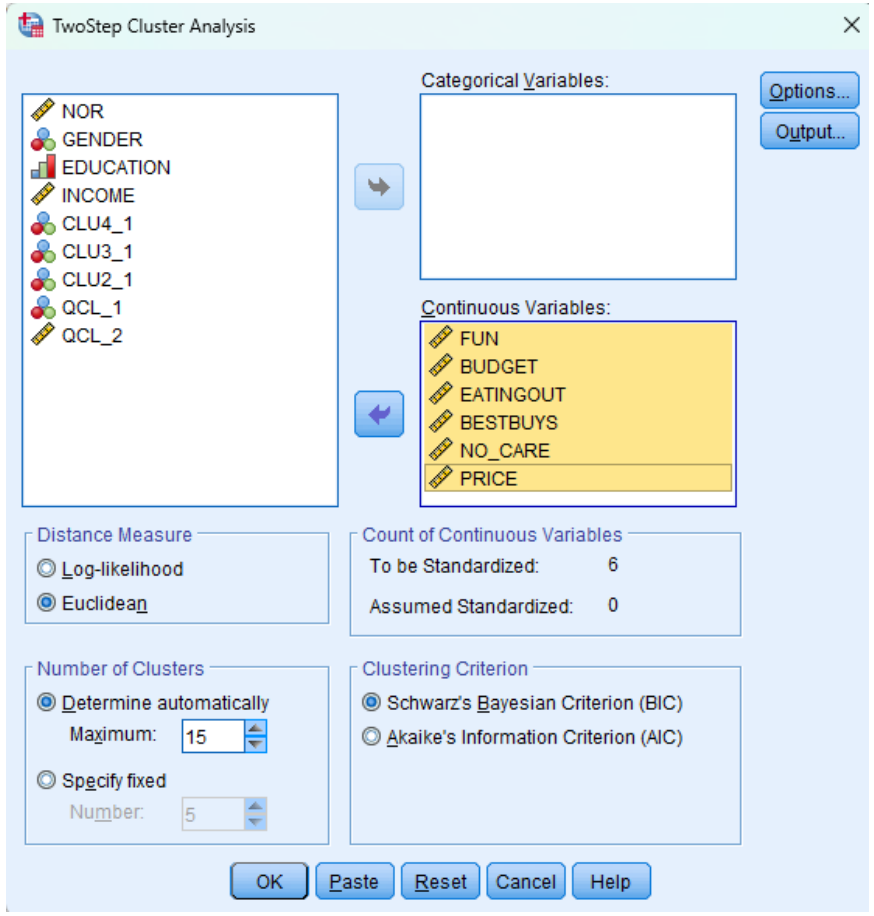


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
			8	0	Number of resp...	None	None	8	Center	Scale	Input
			8	0	Shopping is fun	{1, Extremel...	None	8	Center	Scale	Input
			8	0	Shopping is bu...	{1, Extremel...	None	8	Center	Scale	Input
			8	0	I combine shop...	{1, Extremel...	None	8	Center	Scale	Input
			8	0	I try to get the ...	{1, Extremel...	None	8	Center	Scale	Input
			8	0	I don't care abo...	{1, Extremel...	None	8	Center	Scale	Input
			8	0	You can save a...	{1, Extremel...	None	8	Center	Scale	Input
			8	0	Gender	None	None	8	Center	Nominal	Input
			8	0	Education	None	None	8	Center	Ordinal	Input
			8	0	Net monthly inc...	None	None	8	Center	Scale	Input
			8	0	Ward Method	None	None	10	Right	Nominal	Input
12	CLU3_1	Numeric	8	0	Ward Method	None	None	10	Right	Nominal	Input
13	CLU2_1	Numeric	8	0	Ward Method	None	None	10	Right	Nominal	Input
14	QCL_1	Numeric	8	0	Cluster Number...	None	None	10	Right	Nominal	Input
15	QCL_2	Numeric	20	5	Distance of Ca...	None	None	22	Right	Scale	Input

Two-step cluster analysis

- analysis identifies groupings by running pre-clustering first and then by running hierarchical methods. Because it uses a quick cluster algorithm upfront, it can handle large data sets that would take a long time to compute with hierarchical cluster methods. In this respect, it is a combination of the previous two approaches. Two-step clustering can handle scale and ordinal data in the same model, and it automatically selects the number of clusters.

Two-step output



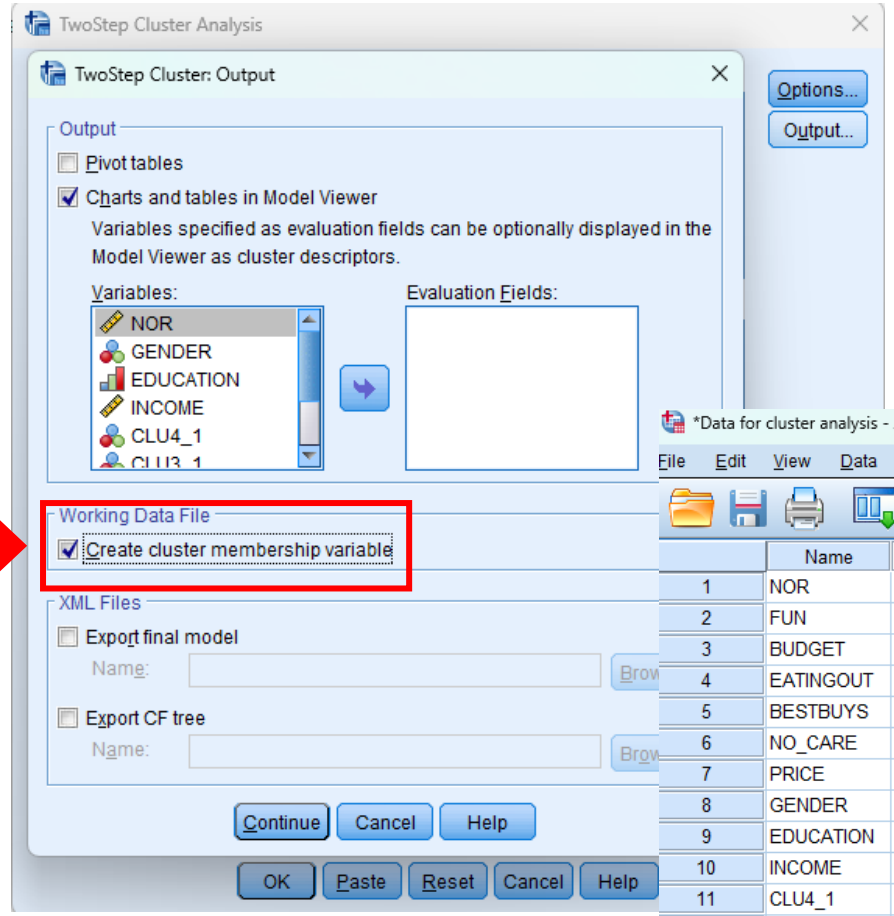
Model Summary

Algorithm	TwoStep
Inputs	6
Clusters	3

Cluster Quality

Create cluster membership variable

Data view



CLU4_1	CLU3_1	CLU2_1	QCL_1	QCL_2	TSC_2086
1	1	1	3	1.41421	3
2	2	2	2	1.32288	1
1	1	1	3	2.54951	3
3	3	2	1	1.40436	2
2	2	2	2	1.84842	1
1	1	1	3	1.22474	3
1	1	1	3	1.50000	3
1	1	1	3	2.12132	3
2	2	2	2	1.75594	1
3	3	2	1	1.14261	2
2	2	2	2	1.04083	1

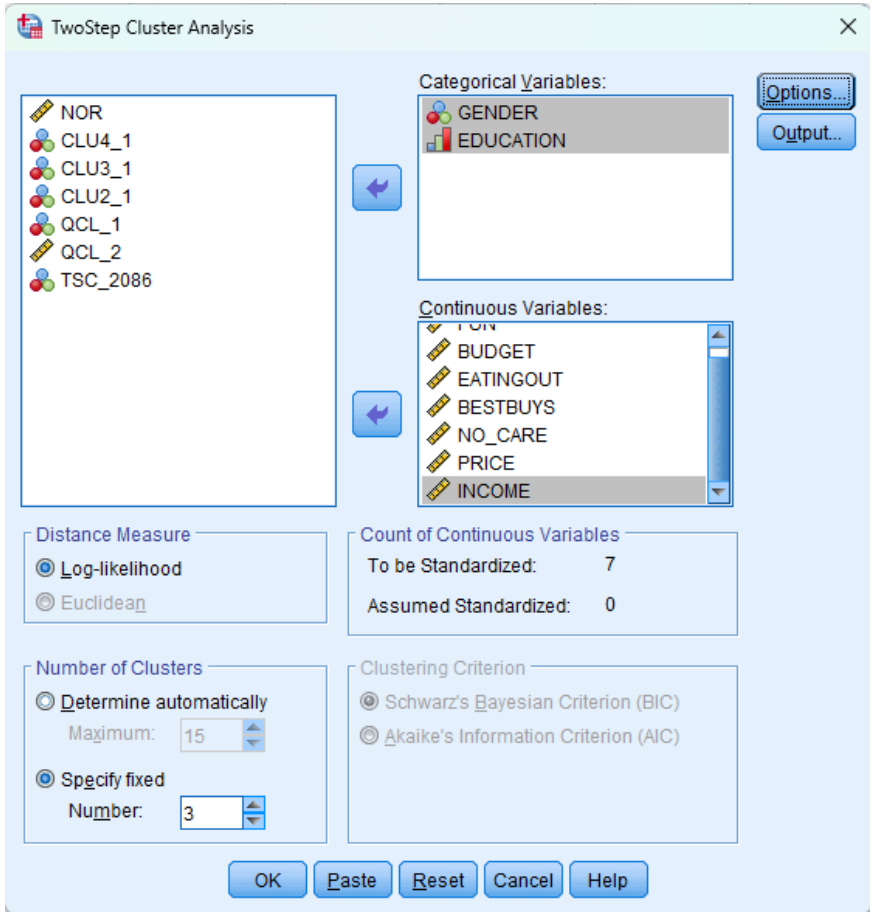
*Data for cluster analysis - Attitudes_Shopping.sav [DataSet2] - IBM SPSS Statisti

File Edit View Data Transform Analyze Graphs Utilities Ex

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	NOR	Numeric	8	0	Number of respondent	None	None	8	Center	Scale	Input
2	FUN	Numeric	8	0	Shopping is fun	{1, Extremel...	None	8	Center	Scale	Input
3	BUDGET	Numeric	8	0	Shopping is bud for your budget	{1, Extremel...	None	8	Center	Scale	Input
4	EATINGOUT	Numeric	8	0	I combine shopping with eating out	{1, Extremel...	None	8	Center	Scale	Input
5	BESTBUYS	Numeric	8	0	I try to get the best buys when shopping	{1, Extremel...	None	8	Center	Scale	Input
6	NO_CARE	Numeric	8	0	I don't care about shopping	{1, Extremel...	None	8	Center	Scale	Input
7	PRICE	Numeric	8	0	You can save a lot of money by comparing prices	{1, Extremel...	None	8	Center	Scale	Input
8	GENDER	Numeric	8	0	Gender	None	None	8	Center	Nominal	Input
9	EDUCATION	Numeric	8	0	Education	None	None	8	Center	Ordinal	Input
10	INCOME	Numeric	8	0	Net monthly income	None	None	8	Center	Scale	Input
11	CLU4_1	Numeric	8	0	Ward Method	None	None	10	Right	Nominal	Input
12	CLU3_1	Numeric	8	0	Ward Method	None	None	10	Right	Nominal	Input
13	CLU2_1	Numeric	8	0	Ward Method	None	None	10	Right	Nominal	Input
14	QCL_1	Numeric	8	0	Cluster Number of Case	None	None	10	Right	Nominal	Input
15	QCL_2	Numeric	20	5	Distance of Case from its Classification Cluster Center	None	None	10	Right	Scale	Input
16	TSC_2086	Numeric	10	0	TwoStep Cluster Number	{-1, Outlier ...	None	8	Right	Nominal	Input

Variable view

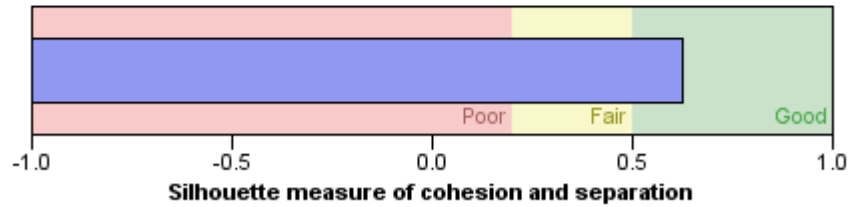
Two-step output เมื่อใส่ตัวแปรรวมกัน



Model Summary

Algorithm	TwoStep
Inputs	9
Clusters	3

Cluster Quality



การจัดกลุ่มจะมีความแตกต่างกันเมื่อใส่ตัวแปรต่างกัน

	PRICE	GENDER	EDUCATION	INCOME	CLU4_1	CLU3_1	CLU2_1	QCL_1	QCL_2	TSC_2086	TSC_2705
1	3	1	1	3000	1	1	1	3	1.41421	3	1
2	4	0	0	2000	2	2	2	2	1.32288	1	2
3	3	1	1	3500	1	1	1	3	2.54951	3	1
4	6	0	0	1500	3	3	2	1	1.40436	2	3
5	4	1	0	2300	2	2	2	2	1.84842	1	2
6	4	1	1	4000	1	1	1	3	1.22474	3	1
7	4	1	1	3800	1	1	1	3	1.50000	3	1
8	4	1	0	4500	1	1	1	3	2.12132	3	1
9	3	1	0	2600	2	2	2	2	1.75594	1	2
10	6	0	0	1600	3	3	2	1	1.14261	2	3
11	3	0	0	2200	2	2	2	2	1.04083	1	2
12	4	1	1	3600	1	1	1	3	1.58114	3	1
13	4	0	0	2400	2	2	2	2	2.59808	1	2
14	7	0	1	1800	3	3	2	1	1.40436	2	3
15	4	1	1	5000	1	1	1	3	2.82843	3	1
16	7	0	0	1650	3	3	2	1	1.62447	2	3
17	5	0	0	2100	1	1	1	3	2.59808	3	3
18	3	0	0	1400	4	3	2	1	3.55512	2	3
19	7	0	0	1600	3	3	2	1	2.15381	2	3
20	2	1	0	2500	2	2	2	2	2.10159	1	2