

제13장 군집분석 (Cluster Analysis)

개요

- 주어진 데이터를 일정한 기준에 집단으로 그룹화하여, 각 집단의 성격을 유사하게 나누는 방법
- n 개의 변수로 이루어진 m 개의 관찰치(레코드)를 그 변수들의 유사도에 따라서 미리 정해진 그룹의 갯수 혹은 동적으로 계산된 그룹의 갯수의 집단(cluster)으로 나누는 방법
- 군집형성의 최대 기준은 그룹내의 데이터의 유사도는 최대한으로 하면서 그룹간의 유사도는 최소한으로 만드는 것

거리 계산의 기본 개념

- 거리의 조건
 - 관찰치 : $X_1 = (X_{11}, X_{12}, \dots, X_{1p})$
 $X_2 = (X_{21}, X_{22}, \dots, X_{2p})$
...
 $X_n = (X_{n1}, X_{n2}, \dots, X_{np})$
 - $d_{ij} = d(X_i, X_j)$: 두 관찰치 X_i, X_j 사이의 거리
 d_{ij} 는 다음의 조건을 만족해야 한다.
 - $d_{ij} \geq 0$ ($i=j$ 이면 $d_{ij} = 0$)
 - $d_{ij} = d_{ji}$
 - $d_{ik} + d_{jk} \geq d_{ij}$ (삼각 부등식)

유사도를 계산법 (연속형)

- 두 관찰치 X_i, X_j 에 대해
 - $X_i = (X_{i1}, X_{i2}, X_{i3}, \dots, X_{in})$
 - $X_j = (X_{j1}, X_{j2}, X_{j3}, \dots, X_{jn})$

- Euclidean Distance

$$d_{ij} = \left[\sum_{k=1}^n (X_{ik} - X_{jk})^2 \right]^{1/2}$$

- Minkowski Distance: 유클리드 거리를 일반화

$$d_{ij} = \left[\sum_{k=1}^n (X_{ik} - X_{jk})^m \right]^{1/m}, \quad m \text{은 실수}$$

- Mahalanobis Distance:

$$d_{ij} = [(X_i - X_j)' S^{-1} (X_i - X_j)]^{1/2}$$

유사도를 계산법 (범주형 변수)

- 범주형 변수
 - $d(X,Y)$: 두 관찰치 X, Y의 불일치 항목수
 - 예) A = (남자, 고졸, 경기), B = (여자, 고졸, 전남), C = (남자, 대졸, 경기)
 - $d(A,B) = 2$, $d(A,C) = 1$, $d(B,C) = 3$

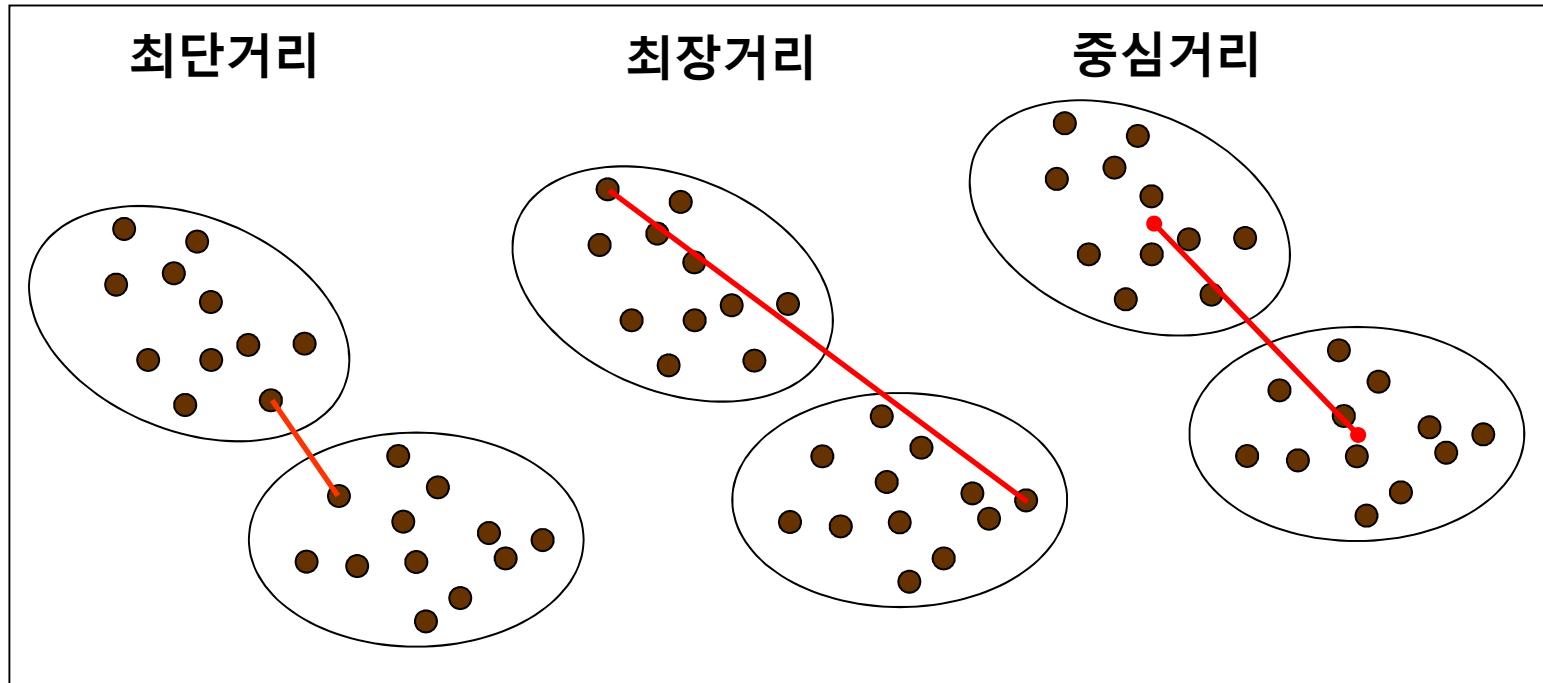
계층적 군집분석

- 병합(agglomeration)과 분할(division)
 - 병합 방법 : 가까운 관측값들끼리 묶어 가는 방법
 - 분할 방법 : 먼 관측값들을 나누어 가는 방법
- 계층적 군집분석에서는 주로 병합 방법을 사용
 - 계층적 군집분석의 결과는 나무구조인 덴드로그램(dendrogram)을 통해 간단하게 나타낼 수 있고,
 - 이를 이용하여 전체 군집들간의 구조적 관계를 쉽게 살펴볼 수 있다.

군집간의 거리를 측정 법 (계층적 군집분석)

- 최단 연결법 (Single Linkage Method)
 - 두 군집의 원소들간 거리를 측정하여 가장 작은 값이 나온 것을 두 군집간 최단 거리로 하는 방법
- 완전 연결법 (Complete Linkage Method)
 - 두 군집의 원소들간 거리를 측정하여 가장 큰 값이 나온 것을 두 군집간 최단 거리로 하는 방법
- 평균 연결법 (Average Linkage Method)
 - 두 군집의 원소들간 거리를 측정하여 평균 값을 두 군집간 최단 거리로 하는 방법
- 중심 연결법 (Centroid Linkage Method)
 - 각 군집의 원소들 평균으로 중심점을 계산한 후 중심점간 거리로 하는 방법
- 미디언 연결법 (Median Linkage Method)
 - 두 군집의 합쳐질 때 새로운 군집의 중심을 이전 군집들의 중심간의 미디언 값으로 하는 방법
- Ward 연결법 (Ward Linkage Method)
 - 군집의 평균에서 군집 내 원소간의 거리를 제곱하여 모두 더한 잔차의 제곱 (error sum of square) 의 증가를 최소화하도록 군집들을 합치는 방법

계층적 군집분석



최단연결법 : 예제

- 다음에 주어진 5개의 관측값에 대한 거리 행렬(유사성 행렬)에 대하여 최단연결법으로 군집을 얻고 덴드로그램으로 나타내보자.

	1	2	3	4	5
1	0				
2	7	0			
3	1	6	0		
4	9	3	8	0	
5	8	5	7	4	0

최단연결법 : 예제

- 거리행렬에서 $d(1,3) = 1$ 이 최소이므로 관측값 1과 3을 묶어 군집 (1,3)을 만든다.
- 거리행렬을 갱신한다.

$$d((1,3),2) = \min\{d(1,2), d(3,2)\} = \min\{7,6\} = 6$$

$$d((1,3),4) = \min\{d(1,4), d(3,4)\} = \min\{9,8\} = 8$$

$$d((1,3),5) = \min\{d(1,5), d(3,5)\} = \min\{8,7\} = 7$$

	(1,3)	2	4	5
(1,3)	0			
2	6	0		
4	8	3	0	
5	7	5	4	0

- $d(2,4) = 3$ 이 최소 \rightarrow 2와 4를 병합

최단연결법 : 예제

- 거리행렬 갱신

$$d((1,3),(2,4)) = \min\{d((1,3),2), d((1,3),4)\} = \min\{6,8\} = 6$$

$$d((2,4),5) = \min\{d(2,5), d(4,5)\} = \min\{5,4\} = 4$$

	(1,3)	(2,4)	5
(1,3)	0		
(2,4)	6	0	
5	7	4	0

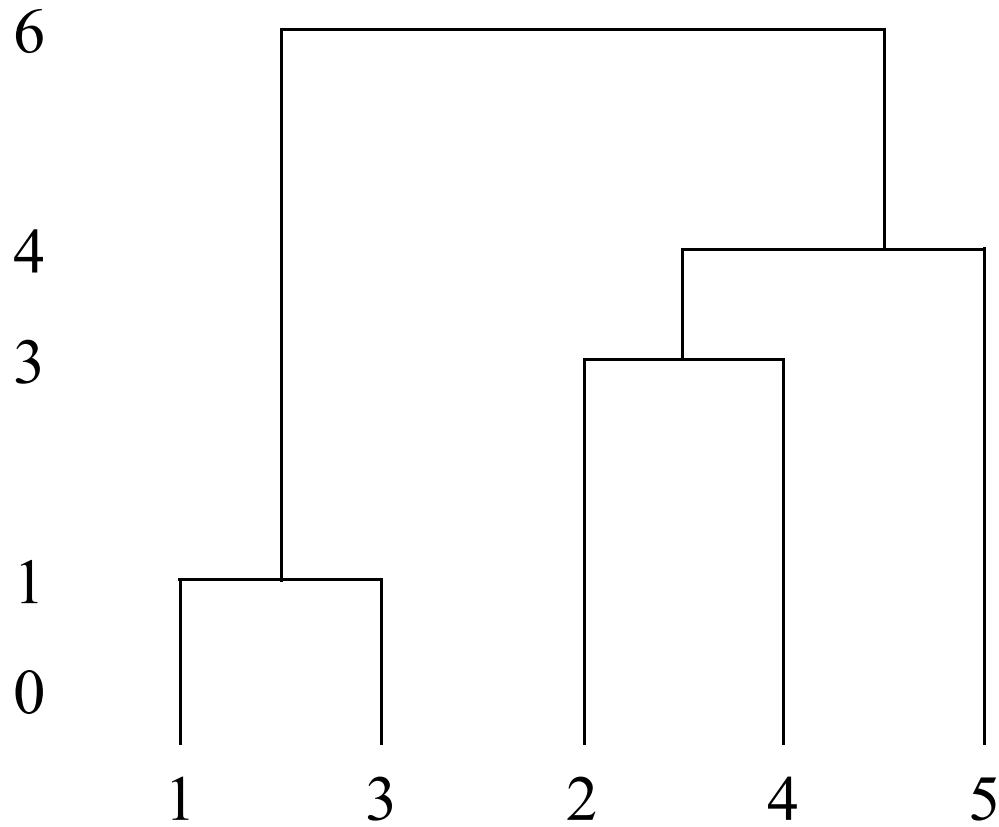
	(1,3)	(2,4,5)
(1,3)	0	
(2,4,5)	6	0

- $d((2,4),5) = 4$ 가 최소 \rightarrow (2,4)와 5를 병합

$$d((1,3),(2,4,5)) = \min\{d((1,3),(2,4)), d((1,3),5)\} = \min\{6,7\} = 6$$

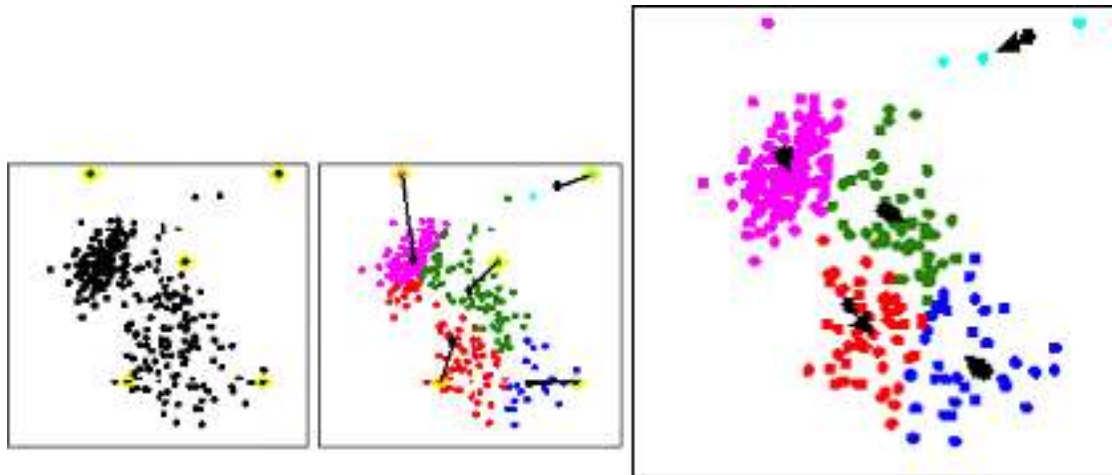
최단연결법 : 예제

- 덴드로그램



비계층적 군집분석

- K-means clustering
 1. 군집의 갯수 k 를 정한다.
 2. 임의의 k 개 점을 택하여 군집의 중심점으로 정한다.
 3. 각 관찰치가 가장 가까운 중심점을 계산하여 특정한 중심점에 가까운 점들은 그 군에 속하는 것으로 간주하다.
 4. 군집내의 점들의 평균을 계산하여 새로운 중심점을 계산한다.
 5. 개체의 할당에 변화가 없을 때까지 3에서 4의 과정을 반복한다.



고려사항

- 가중치 문제
- 군집의 개수 설정 문제
- 단위변환의 문제 : Normalize

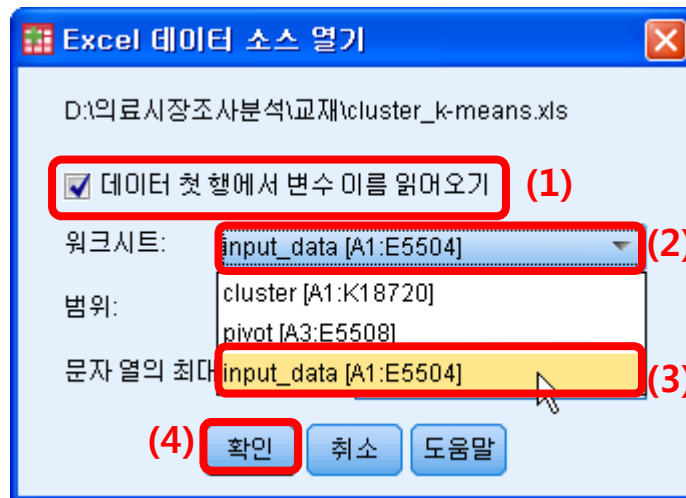
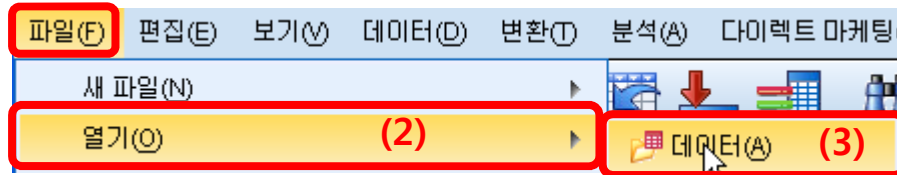
- $$d_{ij} = \left[\sum_{k=1}^p \left\{ \frac{X_{ik} - X_{jk}}{S_k} \right\}^2 \right]^{1/2}$$

실습

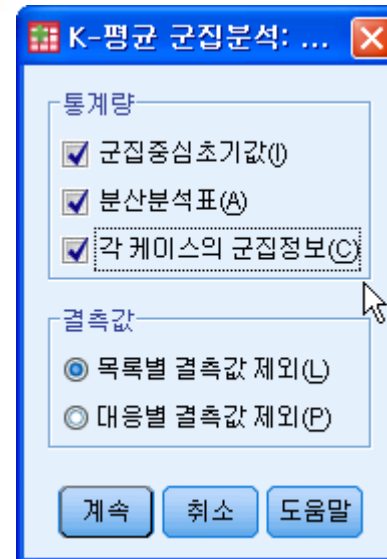
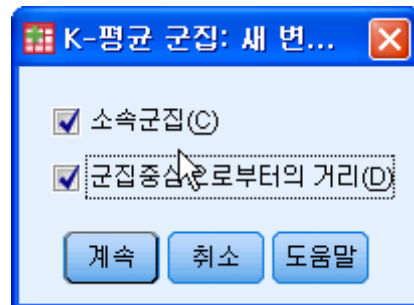
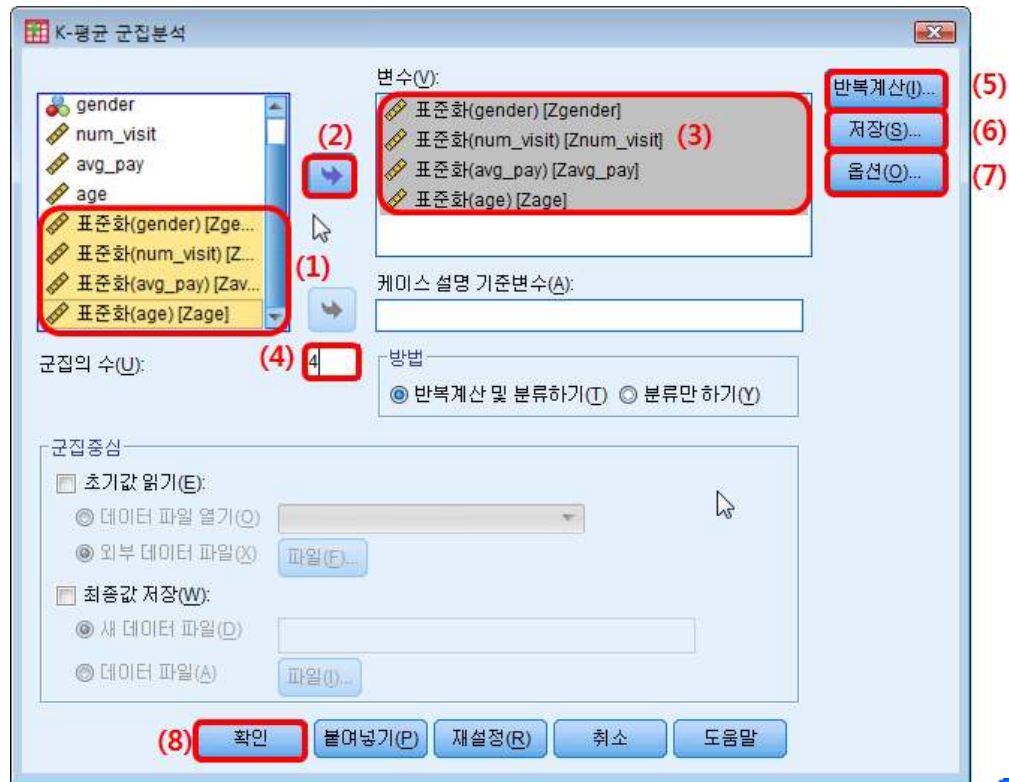
파일명 (사례 수)	변수명	변수설명	척도
cluster_k-means .xls (n=5503)	ID	환자 식별자	명목척도
	gender	성별 (남자 = 0, 여자 = 1)	명목척도
	num_visit	누적 외래방문 횟수	비율척도
	avg_pay	외래방문당 진료비 총액 평균	비율척도
	age	환자 나이	비율척도
cluster_hierarchy .sav (n=49)	ID	환자 식별자	명목척도
	nationality	응답자 국적	명목척도
	num_use	과거 의료관광 서비스 이용횟수	비율척도
	relative_price	자국대비 서비스 이용요금 수준	등간척도
	medical_sat	의료서비스 만족도	등간척도
	tour_sat	관광서비스 만족도	등간척도
	gender	성별 (남자 = 0, 여자 = 1)	명목척도
	age	연령구분	등간척도

실습1 - 치과 외래환자 분석

(1)

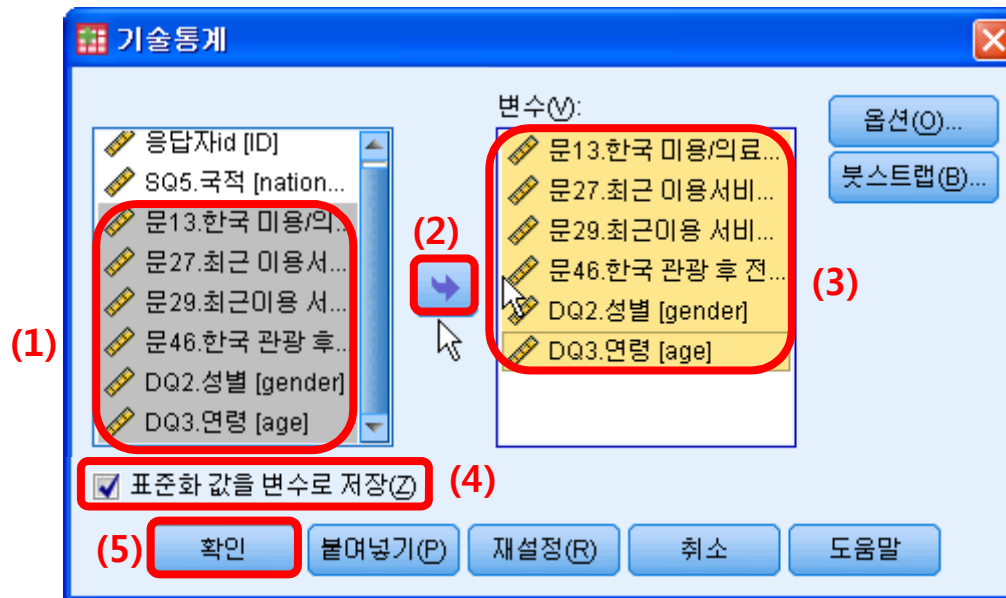
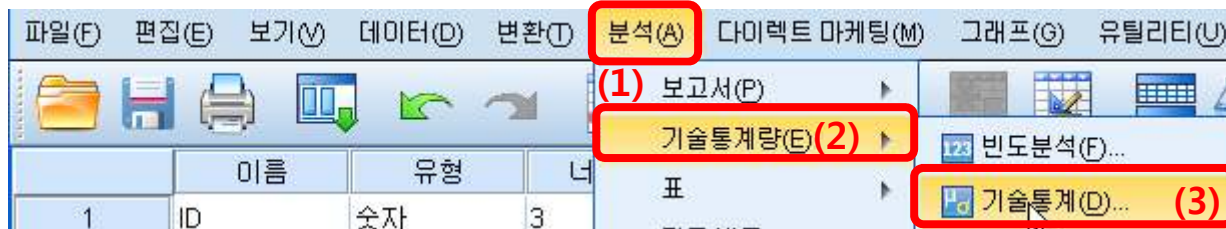


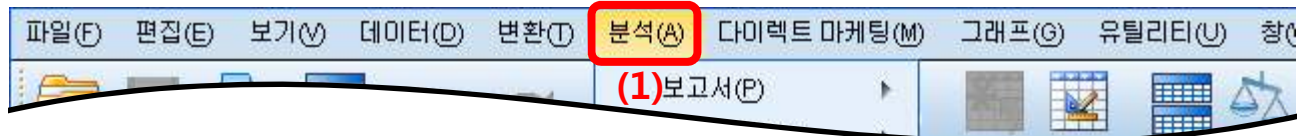




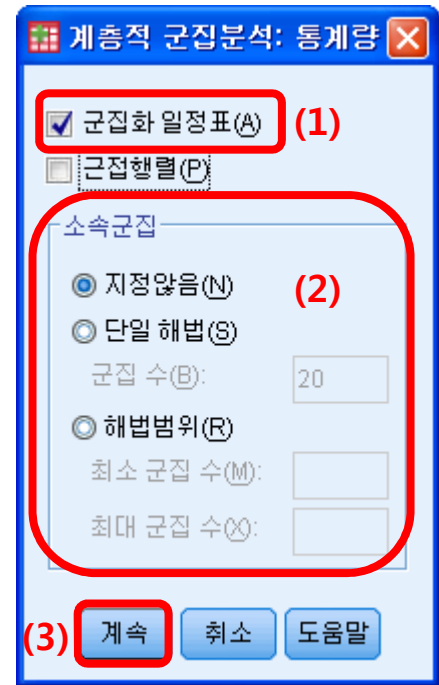
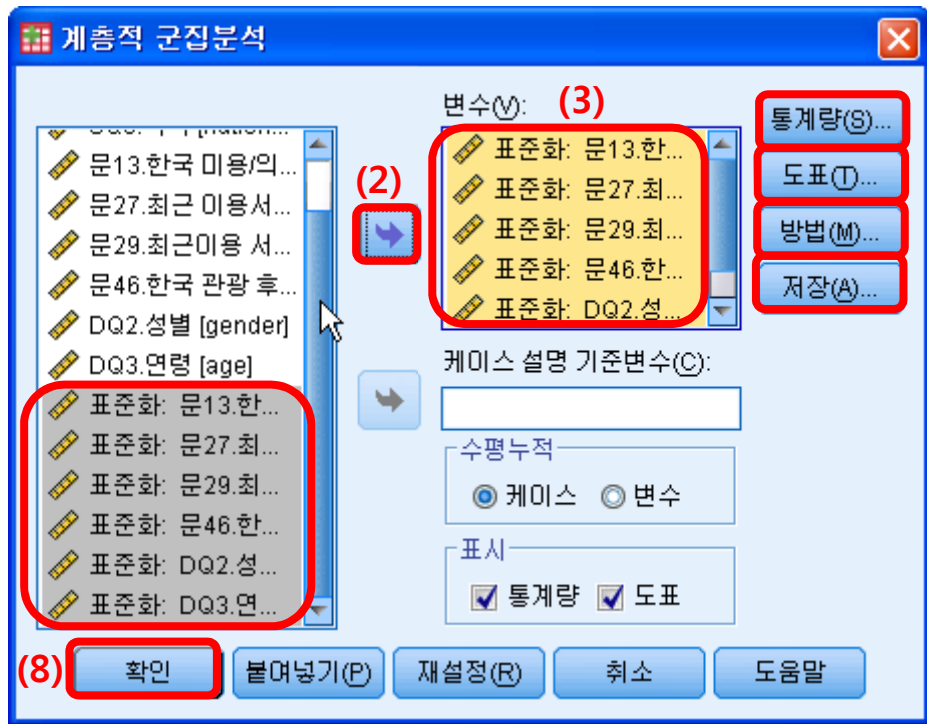
변수	기술 통계량	전체	군집1 n=3087	군집2 n=1878	군집3 n=25	군집4 n=4	군집5 n=125	군집6 n=384
성별	평균	0.49	.47	.51	.52	0.0	.41	.49
	남자 비율	51.5%	53.0%	48.6%	48.0%	100%	59.2%	51.5%
	여자 비율	48.5%	47.0%	51.4%	52.0%	0%	40.8%	48.5%
누적 외래 방문 횟수	평균	3.4	2.45	2.63	8.44	5.50	8.65	12.72
	표준 편차	0.50	1.62	1.73	7.63	2.65	5.35	4.50
	최대	39	9	9	27	8	25	39
	최소	1	1	1	1	2	1	8
외래 방문당 진료비 총액 평균	평균	52	46	37	610	1,149	260	56
	표준 편차	66	29	24	107	141	65	40
	최대	1298	205	189	859	1,298	442	235
	최소	0	0	0	445	969	159	12
환자 나이	평균	38.3	24.3	59.4	49.6	61.2	42.8	45.1
	표준 편차	20.0	9.7	11.7	15.7	16.3	16.6	18.0
	최대	91	0.0	91.0	15.7	43.0	88.0	87.1
	최소	0	44.0	41.0	73.0	77.6	10.0	8.4

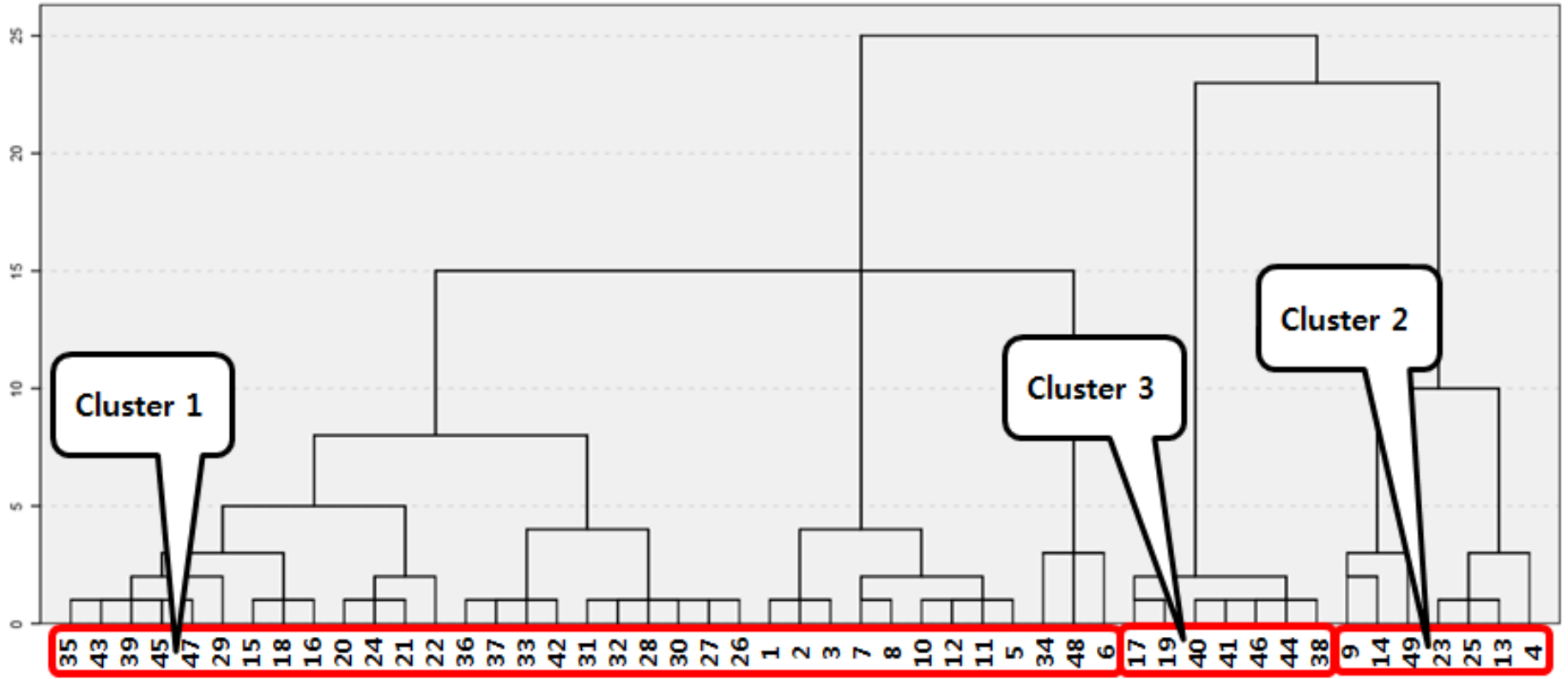
실습2 - 의료관광 응답자 분석





9	Znum_use	숫자	11	전송...
10	Zrelative_price	숫자	11	분류분석(Y) (2)
11	Zmedical_sat	숫자	11	차원 감소(D)
12	Ztour_sat	숫자	11	척도(A)
13	Zgender	숫자	11	계층적 군집분석(H) (3)





변수	기술 통계량	전체 n=49	군집 1 n=35	군집 2 n=7	군집 3 n=7
성별	남자 (비율)	8 (16.3%)	0 (0%)	1 (14.3%)	7 (100%)
	여자 (비율)	44 (83.7%)	35 (100%)	6 (85.7%)	0 (0%)
국적	일본 (비율)	24 (49.0%)	24 (68.6%)	0 (0%)	0 (0%)
	미국 (비율)	19 (12.2%)	10 (28.6%)	2 (28.6%)	7 (100%)
	기타 (비율)	1 (2.9%)	1 (2.9%)	4 (57.1%)	0 (0%)
연령대	20~29세	6 (12.2%)	0 (0.0%)	0 (0.0%)	6 (17.1%)
	30~39세	14 (28.6%)	3 (42.9%)	0 (0.0%)	11 (31.4%)
	40~49세	11 (22.4%)	1 (14.3%)	1 (14.3%)	9 (25.7%)
	50~59세	10 (20.4%)	1 (14.3%)	3 (42.9%)	6 (17.1%)
	60세 이상	8 (16.3%)	2 (28.6%)	3 (42.9%)	3 (8.6%)
미용/의료서 비스 이용횟수	평균	2.51	2.97	1.14	1.57
	표준 편차	2.599	2.935	.378	.787
	최대	14	14	2	3
	최소	1	1	1	1

자국대비 한국 요금	20% 이하	4	4	0	0
	21~40%	7	6	0	1
	41~60%	9	4	0	5
	61~80%	8	8	0	0
	81~100%	11	9	1	1
	101~120%	5	4	1	0
	141~160%	2	0	2	0
	그 이상	3	0	3	0
최근 이용 서비스 만족도	평균	4.39	4.31	4.14	5
	표준 편차	0.786	.796	.900	0
	최대	5	5	5	5
	최소	3	3	3	5
관광 만족도	평균	4.39	4.40	4.00	4.71
	표준 편차	0.671	.695	.577	.488
	최대	5	5	5	5
	최소	3	3	3	4