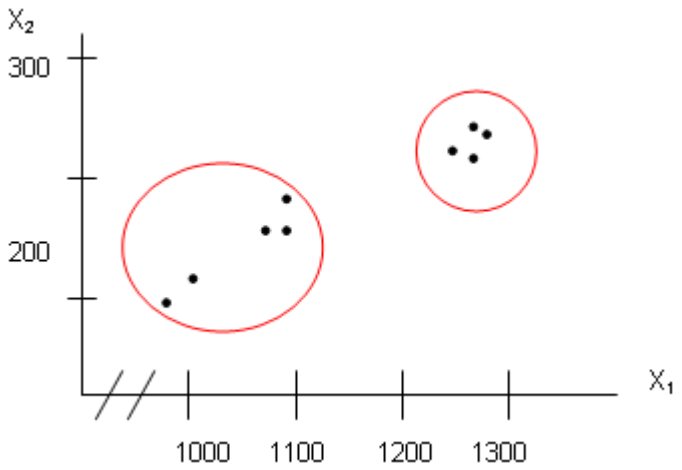
Chapter 5. 군집분석

마케팅 담당자가 자사 고객들을 분류하기 위하여 나이, 학력, 소득, 결혼 상태, 자녀 수, 직업 등에 대한 정보를 수집하였다. 이를 이용하여 고객들을 집단(cluster)으로 분류하고자 할 때 사용되는 다변량 분석 방법이 군집분석 (Clustering Analysis)이다. 판별분석과 군집분석의 다른 점은 판별분석은 조사된 데이터에 개체의 집단 변수가 이미 포함되어 있으나 군집분석은 개체들에 대해 측정된 변수에 의해 집단을 분류하게 되므로 집단의 개수와 집단의 종류(이름)는 분류 후 정해지게 된다. 즉 군집분석은 개체들이 분석 전에는 어떤 그룹에 속하는지 알려져 있지 않다. 군집분석은 grouping 혹은 classification 이라 불리기도 한다.

개체를 분류한다? 아래 산점도와 같이 측정 변수가 2 개라면 거리가 가까운 개체들끼리 묶으면 될 것이다. 12 개의 개체가 아마 2 개 군집(집단)으로 분류될 것이다. 두 개체간의 Euclidean 거리를 계산(측정)하고 거리가 가까운(유사성이 높음) 개체끼리 묶으면 된다.



5.1 군집분석 개념

5.1.1 유사성

개체들의 유사성(거리)은 다음에 의해 측정한다.

▣Euclidean 거리

두 개체 사이의 우사 정도를 거리로 표현할 수 있다. 거리가 멀면 유사성(similarity)이 떨어진다. r번째 개체와 s번째 개체의 Euclidean 거리 식은 다음과 같다.

$$d_{rs} = [(\underline{x}_r - \underline{x}_s)'(\underline{x}_r - \underline{x}_s)]^{1/2}$$

▣표준화 Euclidean 거리

개체들에 대해 측정된 변수들이 단위가 다르거나 분산이 다를 경우 변수를 표준화한 후 거리를 구하는 것이 더 적절하다. 다음 식은 r 번째 개체와 s 번째 개체의 표준화 Euclidean 거리이다.

$$d_{rs} = [(\underline{z}_r - \underline{z}_s)'(\underline{z}_r - \underline{z}_s)]^{1/2}$$

▣Mahalanobis 거리

다음 식이 r번째 개체와 s번째 개체의 Mahalanobis 거리이고 Σ는 within 군집 분산-공분산 행렬 추정치이다. 거리를 이와 같이 정의하는데 문제점은 개체를 다 분류하기 전에는 Σ의 추정치를 구할 수 없다는 것이다.

$$d_{rs} = [(\underline{x}_r - \underline{x}_s)'\Sigma^{-1}(\underline{x}_r - \underline{x}_s)]^{1/2}$$

5.1.2 군집분석 방법

▣비계층적 방법

군집의 중심이 되는 seed 점들 집합을 선택하여 그 seed 점과 유사성이 높은(거리가 가까운) 개체들을 묶는(그룹화) 방법이다. 이 방법은 다음 3 가지 문제점을 갖고 있다. 1)사전에 군집(그룹) 수에 대한 예상이 필요하다. 2)개체 분류는 처음 선정한 seed 점들에 의해

영향을 많이 받고 분석자 마다 분류가 다를 가능성이 있다. 3)군집의 수와 seed 값의 위치의 결합 조건이 너무 많아 계산이 분류를 위한 계산이 용이하지 않다.

▣계층적 방법

유사성이 가까운 순서대로 개체들을 묶어(군집화) 가는 방법으로 single-linkage clustering 방법이 이 방법 중 가장 효율적이다. Neighbor Method 은 single-linkage clustering 방법 중 하나로 다음 순서에 의해 개체를 분류한다.

(1)처음에는 개체의 수(n)만큼의 군집이 있다. 예를 들어 개체 6 개가 있고 다음은 각 개체 간 Euclidean 거리(유사성)를 계산한 표이다. 처음에는 군집이 6 개이다.

	1	2	3	4
1		0.1	0.7	0.2
2			0.4	0.6
3				0.3
4				

두 개체간의 거리이므로 대각 원소는 동일하다.

(2)유사성(거리)이 가장 가까운 개체를 군집으로 묶는다. 예제에서는 (3.5)가 묶인다.

	(1, 2)	3	4
(1, 2)		?	?
3			0.3
4			

?: 어떤 값으로 개체 집단(1, 2)와 개체)의 유사성(거리)을 측정할 것인가?

(3)개체가 군집으로 묶이면 개체와 새로 만들어진 군집과의 유사성을 계산한다. 군집과 군집(혹은 개체)의 유사성(거리)을 측정하는 방법은 다음 5 가지가 있다.

①Nearest neighbor: 두 군집의 각 개체 중 가장 가까이 있는 개체의 거리(유사성)

②Furthest neighbor: 두 군집의 각 개체 중 가장 멀리 있는 개체의 거리

③Centroid neighbor: 군집의 평균 간의 거리

④Average neighbor: 한 군집의 개체와 다른 군집 개체들의 각 거리 평균

⑤Ward's minimum variance: 군집의 평균간 거리를 각 군집의 개체 개수의 역의 합으로 나눈 제곱근을 구한 거리이다.

Nearest, Furthest, Centroid neighbor, Average neighbor, Ward's minimum variance 중 어떤 방법을 사용하는 것이 좋은가? Nearest 방법은 개체간의 거리가 가까워 개체를 묶는 경향이 있어 군집의 수가 줄어들고 Furthest 는 군집간 거리를 최소화 하는 경향이 있어 개체수가 적은 군집을 얻게 한다. 그러므로 각 방법의 장단점이 있으므로 2-3 개 방법을 사용하여 개체의 군집화가 보다 잘되는 방법을 선택하는 것이 좋다. 가장 많이 사용하는 방법은 Average neighbor 방법이다.

(4)다음은 Nearest neighbor 방법에 의해 개체를 군집화 하는 과정이다.

1 과 3 의 거리는 0.7, 2 와 3 의 거리는 0.4 이므로 (1, 2)와 3 의 거리는 0.4 가 된다. 1 과 4 의 거리는 0.2 이고 2 와 4 의 거리는 0.6 이므로 작은 거리 0.2 가 (1, 2)와 4 의 거리이다.

	(1, 2)	3	4
(1, 2)		0.4	0.2
3			0.3
4			

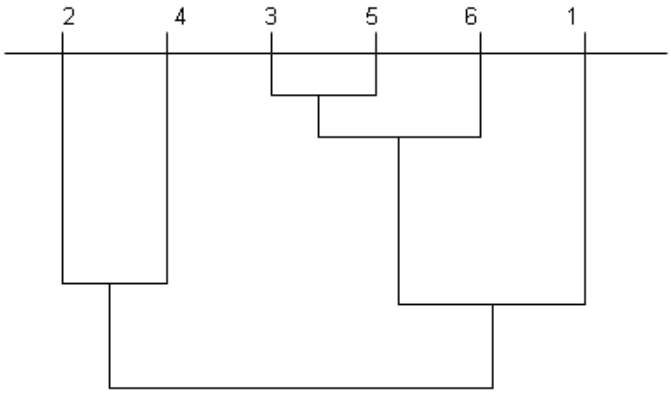
(1, 2)의 거리는 0.4 이고 3 과 4 의 거리는 0.3 이므로 (1, 2, 4)와 3 의 거리는 0.3 이 된다.

	(1, 2, 4)	3
(1, 2, 4)		0.3
3		

5.1.3 군집 개수

군집의 개수를 몇 개로 하면 좋은가? 그래프적 방법으로는 나무 다이어그램(Tree diagram)이 있고 검정 통계량을 이용하는 방법으로는 Hotel ling's T^2 검정이나 Cubic Clustering Criterion 방법을 이용하면 된다.

▣계층적 나무 다이어그램



위의 그림은 나무 그림(Tree Diagram)이라 하여 선의 길이는 개체간, 개체와 군집간, 군집간 유사성(거리)이다. 이 다이어그램(diagram)에 의하면 2 개의 군집 (2, 4), (3, 5, 6, 1)으로 분류하는 것이 타당해 보인다.

▣Pseudo Hotel ling's T^2 검정

Hotel ling's T^2 검정 통계량은 두 집단 다변량 평균의 차이를 보는 통계량이다. 이를 군집분석에 이용하는데..... 이와 유사 개념의 검정 통계량을 이용하여 개체의 군집간 평균의 차이가 유의하지 않으면 두 군집을 합치고 유의하면 군집 그대로 유지하는 방법이다.

▣CCC

Searle(1983)이 제안한 방법으로 군집의 개수와 CCC(Cubic Clustering Criterion)의 산점도를 그려 CCC 의 값이 3 이상이고 최대 값인 경우 그 때의 군집의 개수가 적당하다. 이용 방법은 예제에서 살펴보기로 한다.

5.1.4 판별분석과 비교

	판별분석(Discriminant Analysis)	군집분석(Clustering Analysis)
분석 초기	개체들은 이미 분류되어 있다. 	개체들을 측정 변수에 의해 분류한다. 
	판별 변수 $(X_1, X_2, ..., X_P)$ 분류 변수	
목적	새로운 개체를 분류 ▶개체를 잘 판별할 수 있는 판별 변수 선택이 관건이다.	위의 개체들을 분류 ▶개체들의 특성을 나타내는 변수들을 선택하는 것이 관건
분석 순서	(1)판별분석 방법 선택 →오분류가 적은 방법 사용 •Fisher method (판별 변수 선택 방법 이용, 유의 수준을 다소 높게 설정) •K Nearest Discriminant Analysis •Logistic Regression 판별분석(변수 선택 방법 사용, 유의수준 다소 높게 설정) (2)개체 분류 경향을 파악하기 위하여 판별 변수들에 산점도(by 그룹)를 그린다. 판별 변수가 2개 이상이면 주성분분석을 이용하여 산점도를 그리면 된다. (3)최종적으로 구해진 판별식에 의해 새로운 개체를 (□△○) 분류한다.	(1)개체 분류 방법을 선택한다. •Nearest neighbor •Furthest neighbor •Centroid neighbor •Average neighbor •Ward's minimum variance (2)군집의 개수를 정한다. •CCC •Pseudo Hotel ling's T^2 •Tree Diagram (3)개체 분류가 잘 되었는지 알아보기 위하여 산점도를 그린다. 변수가 3개 이상인 경우는 주성분분석을 이용하여 산점도 그린다. 군집 결과는 개체 분류 방법과 군집 개수에 의해 결정된다. (4)각 군집에 적절한 이름을 붙인다. 

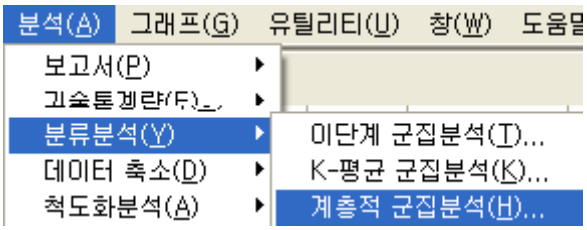
5.2 군집분석하기

56 개 피자 제품에 대해 MOIS(수분 함유 X1), PROT(단백질 함유량 X2), FAT(지방 함유량 X3), ASH(ash 함유량 X4), SODIUM(나트륨 함유량 X5), CARB(탄수화물 함유량 X6), CAL(칼로리 X7)를 조사하였다. 이를 이용하여 56 개 피자 제품을 분류하여 보자.

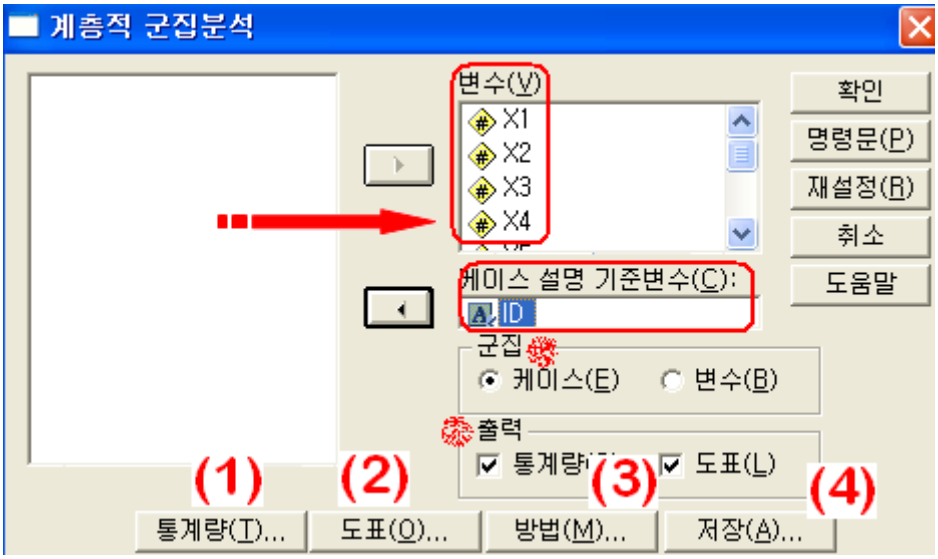
[Applied Multivariate Methods for Data Analysts, Dallas E. Johnson, 1998] ■PIZZA.SAV■

ID	X1	X2	X3	X4	X5	X6	X7
14025	28.35	19.99	45.78	5.08	1.63	.80	4.95
14164	28.70	20.00	45.12	4.93	1.56	1.25	4.91
14154	30.91	19.65	42.45	4.81	1.65	2.81	4.72
24082	31.02	19.05	42.29	5.27	1.71	2.37	4.66
24138	29.62	21.10	43.37	5.05	1.69	.86	4.78

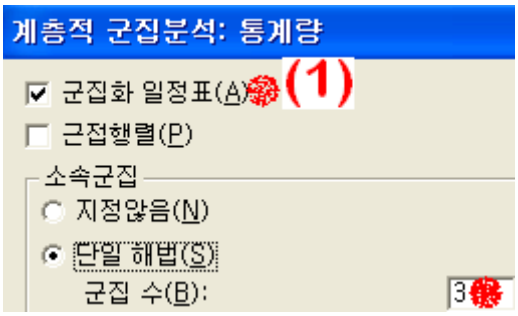
5.2.1 계층적 군집분석



군집에 사용될 변수를 지정한다. 케이스 구별 변수를 지정한다. 아래 1, 2, 3, 4 옵션은 필요에 따라 지정하면 된다.



군집화 일정은 가장 가까운 개체부터 묶여지는 과정을 보여준다. 3 번째는 군집 1 과 군집 2 가 묶였다는 것을 의미한다. 단일 해법에서 군집의 수를 3 으로 했으므로 각 개체의 소속 군집이 1, 2, 3 중 하나로 출력된다. "소속 군집"을 굳이 선택할 필요는 없다. Why? 필요하지 않다.



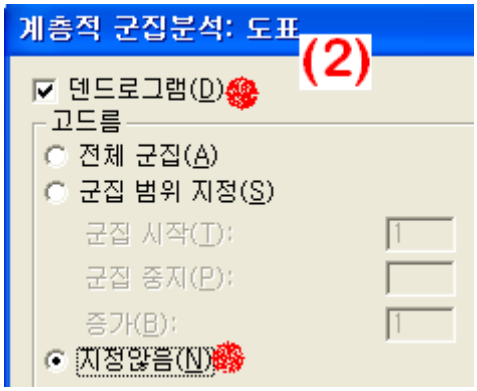
군집화 일정표

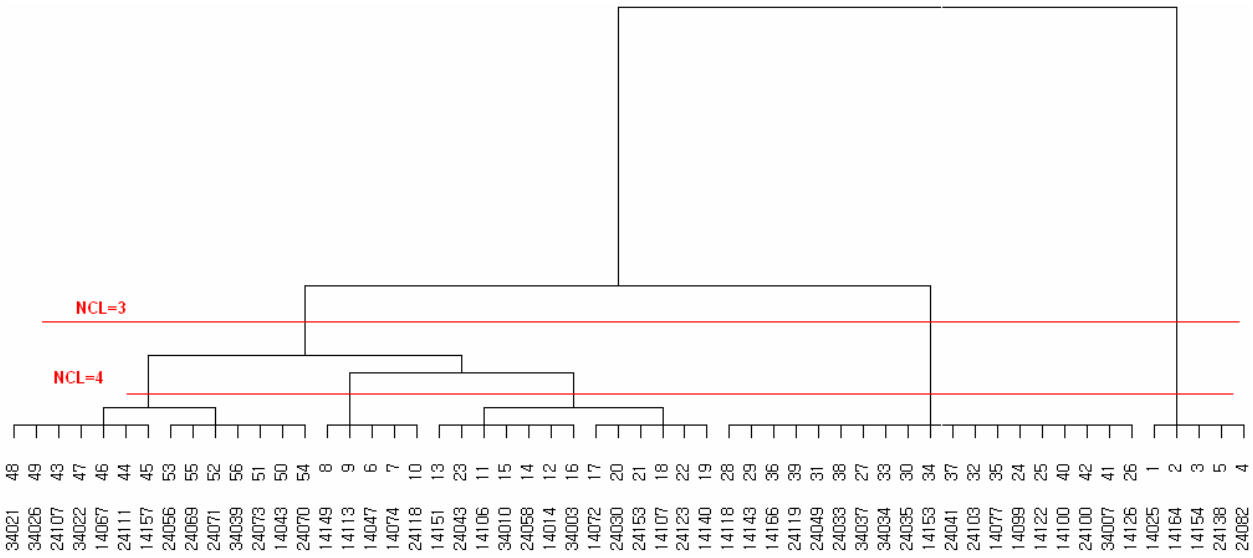
단계	결합 군집		계수	처음 나타나는 군집의 단계	
	군집 1	군집 2		군집 1	군집 2
1	48	49	.006	0	0
2	43	47	.013	0	0
3	43	48	.016	2	1
4	17	20	.019	0	
5	43	46	.019	3	0

소속군집

케이스	3 군집
1:14025	1
2:14164	1
3:14154	1
4:24082	1
5:24138	1
6:14047	2
7:14074	2
8:14149	2

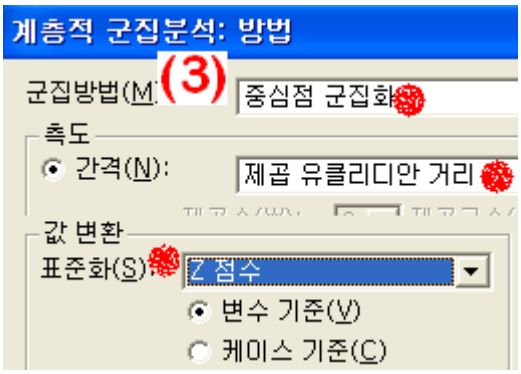
나무 다이어그램(Dendrogram)을 출력하라는 옵션이다. 고드름 도표는 보기 복잡하고 활용 면에서 떨어지므로 출력하지 않는 것이 좋다. SPSS 출력 창에는 덴드로그램이 수직으로 출력된다. 보기 편하기 하기 위하여 돌려 놓은 것이다. 군집을 4 개로 하면 어떨까? 불행히도 SPSS 에서는 군집 개수 결정에 참고가 되는 통계량이 출력되지 않는다. 사실 통계량은 큰 도움이 되지 않는다.

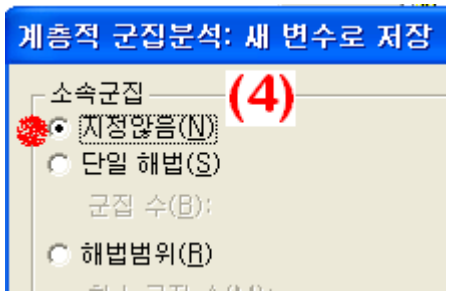




군집 방법은 중심점 군집화(Centroid clustering) 방법을 사용하였고 유사성(거리)는 Euclidian 제곱 거리를 사용하였다. 어느 방법이 최선? 아무도 모른다. 값의 표준화 점수를 사용한 이유는 군집에 사용되는 변수의 단위가 다르기 때문이다.

덴드로그램을 보고 군집의 개수를 결정하기 전까지는 소속 군집을 출력하지 않는 것이 좋다.





덴드로그램을 참조한 결과 군집의 개수는 4 개였다. 이제 “소속군집”에서 “단일해법”에서 군집의 수를 4 로 하여 군집분석을 재실시하면 아래와 같이 데이터 내에 군집 변수가 저장된다.

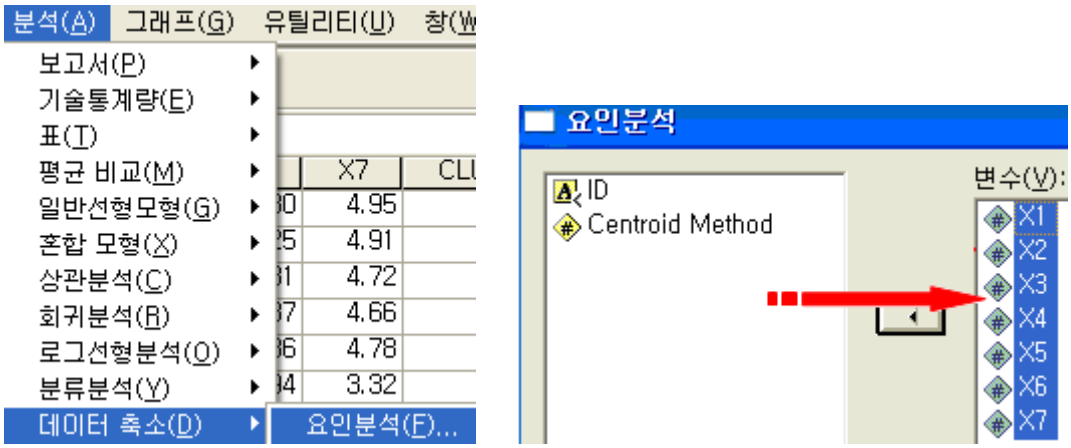
ID	X1	X2	X3	X4	X5	X6	X7	CLU4_1
14025	28.35	19.99	45.78	5.08	1.63	.80	4.95	1
14164	28.70	20.00	45.12	4.93	1.56	1.25	4.91	1
14154	30.91	19.65	42.45	4.81	1.65	2.81	4.72	1
24082	31.02	19.05	42.29	5.27	1.71	2.37	4.66	1
24138	29.62	21.10	43.37	5.05	1.69	.86	4.78	1
14047	49.99	13.35	29.20	3.52	1.05	3.94	3.32	2
14074	50.72	12.93	29.88	3.60	1.03	2.87	3.32	2
14149	54.96	14.26	22.99	3.19	.90	4.60	2.82	2
14113	54.12	14.06	24.95	3.14	.82	3.73	2.96	2

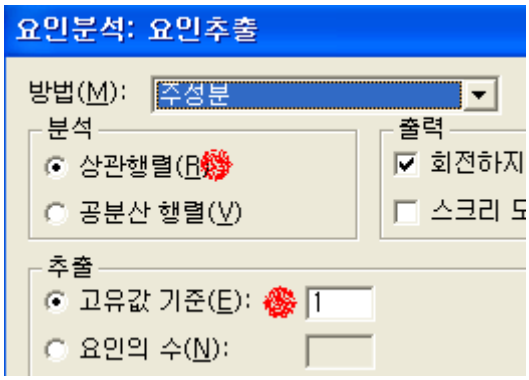
5.2.2 군집 이름 부여하기

개체 군집 결과를 위의 출력 형태보다는 그래프로 보면 이름을 붙이기가 더 쉬울 것이다. 개체 군집이 나타날 수 있도록 산점도를 그려 보자. 그런데 불행히도 각 피자에 대해 측정 변수가 7 개(MOIS, PROT, FAT, ASH, SODIUM, CARB, CAL)이므로 하나의 산점도로는 표현할 수 없다. 그리고 군집분석에서는 분류 변수(군집분석에 이용되는 변수)를 선택한다는 것은 전혀 의미가 없다.

유용한 그래프 산점도를 어떻게 그리지? 변수가 2 개라면 몰라도... 아 어떻게 하지. 고민하지 말자. 우리는 변수의 개수를 축약하는 방법인 주성분분석을 알고 있다. 주성분분석에 의해 분류 변수를 축약하고 주성분분석으로 산점도를 그리자. 주성분 개수가 2 개 이하이면 더 말할 나위 없이 좋지만 3 개라도 산점도 2 개만 그리면 되니 별 문제는 없다.

주성분 2 개만 구하여 군집 결과를 출력해보자.





성분점수 계수 크기를 이용하여 제일 주성분은 영양소 함유량 변수(?), 제이 주성분은 칼로리 변수라 이름을 부여하였다.

설명된 총분산

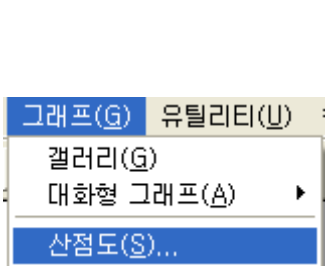
성분	초기 고유값		
	전체	% 분산	% 누적
1	3,909	55,845	55,845
2	2,523	36,048	91,893
3	,442	6,321	98,214
4	,098	1,405	99,619
5	,027	,380	99,999
6	6,327E-05	,001	100,000
7	8,871E-06	,000	100,000

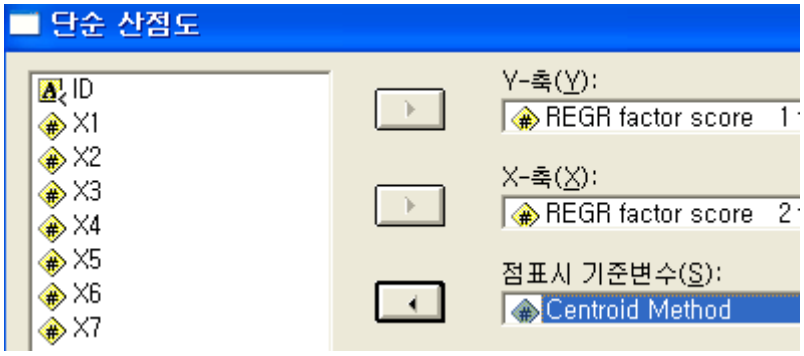
추출 방법: 주성분 분석.

성분점수 계수행렬

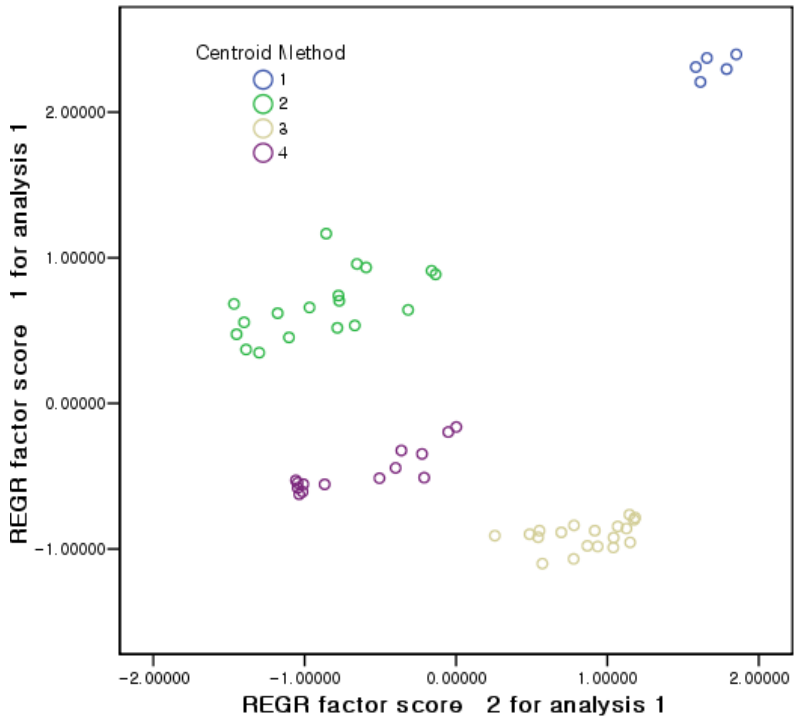
	성분	
	1	2
X1	,037	-,372
X2	,190	-,181
X3	,222	,177
X4	,246	-,068
X5	,223	,143
X6	-,218	,201
X7	,106	,358

제일 주성분, 제이 주성분의 산점도를 그려보자. 산점도의 점 표시 변수는 군집 변수(CLU*_1)를 설정하였다.

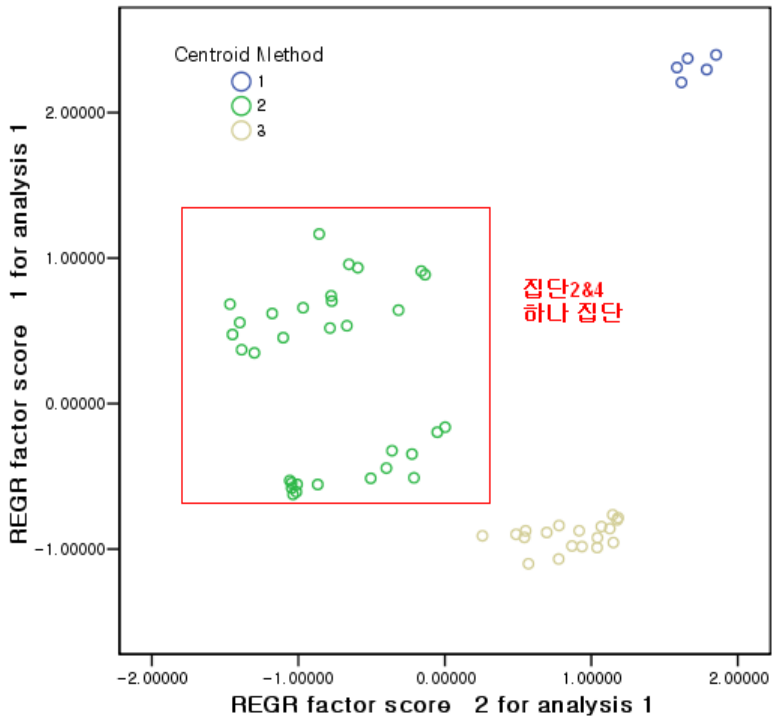




군집이 명확히 구별된다. 집단 1 은 영양소 함유량 높고 칼로리 높은 피자이다. 제일 주성분에 의해 군집이 나누어진다. 제이 주성분은 집단 2 와 4 를 구별하지 못한다.



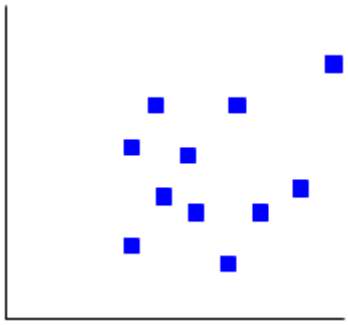
만약 군집의 개수를 3 으로 하여 산점도를 구하면 다음과 같다.



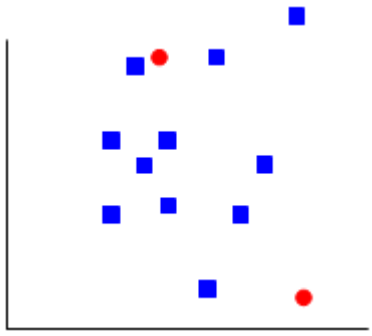
5.2.3 Faster Cluster 군집분석

Faster clustering 방법은 앞 절에서 살펴보았던 계층적 군집(hierarchical clustering) 방법과는 [유사성(거리)이 가까운 개체들을 차례로 군집으로 묶어가는 방법] 달리 비계층적 군집(non-hierarchical clustering) 방법이다. 우선 seed 를 정하고 이 seed 에 가까운 개체들을 군집으로 묶는다. 그러므로 군집의 개수를 분석 전에 정해야 하며(number of clusters) 군집의 크기(size: 이는 radius 로 설정)를 정해 주어야 한다. 비계층적 군집 방법의 순서는 다음과 같으며 SAS 에서는 FASTCLUS 에서 이 방법으로 군집분석할 수 있다.

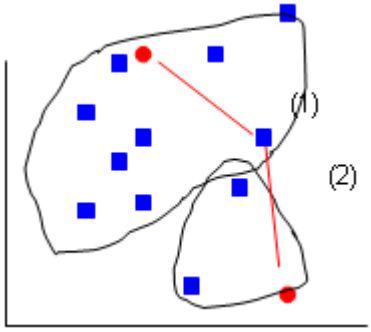
Faster Cluster 방법을 이해하기 위하여 예를 들어 설명하기로 한다.



(STEP1)군집 seeds 가 선택된다. 초기 seed 의 개수는 분석자가 정해 주는 개수대로 된다. **MAXCLUSETRS=2** 옵션을 사용했을 경우 초기 SEED 개수는 2 개이다.

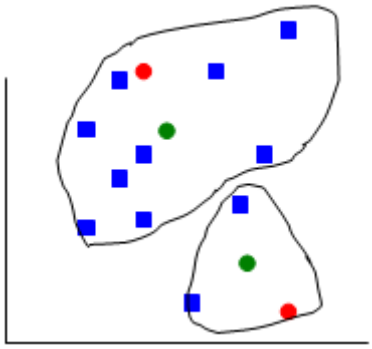


(STEP2)개체들을 가장 가까운 군집 seeds 에 묶는다. 만약 이 경우 DRIFT 옵션을 쓰면 seed 가 임시 군집의 평균으로 옮겨가며 개체를 묶는다.

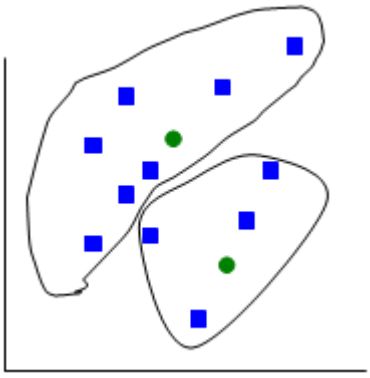


(1)이 (2)보다 거리가 가깝다.

(STEP3)일단 개체 군집이 끝나면 군집 개체의 평균을 SEED 로 하여 개체를 다시 분류한다.



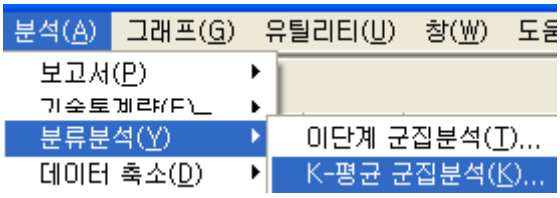
(STEP4)군집이 끝나면 STEP 2)-STEP3)을 반복한다. 다음 STEP 에서는 SEED 가 빨강에서 초록색으로 옮겨간다. 그리고 위의 STEP 을 반복한다. MAXITER 옵션이 없으면 개체가 군집 분류가 이전과 같을 때까지 반복한다. MAXITER=3 이라 하면 위의 STEP 을 3 번 반복한다.



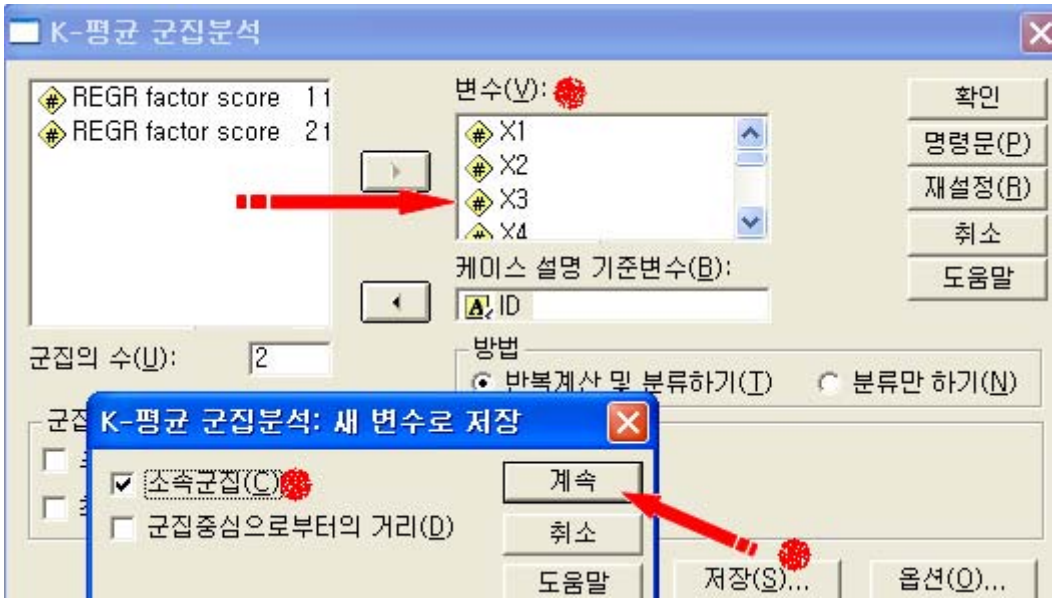
Non-hierarchical clustering 방법의 또 하나의 결점은 자료 입력 순서에 따라 군집이 달라지므로 RANDOM 이라는 옵션을 사용해 자료 입력 순서를 바꾸어 가며 군집하게 한다.

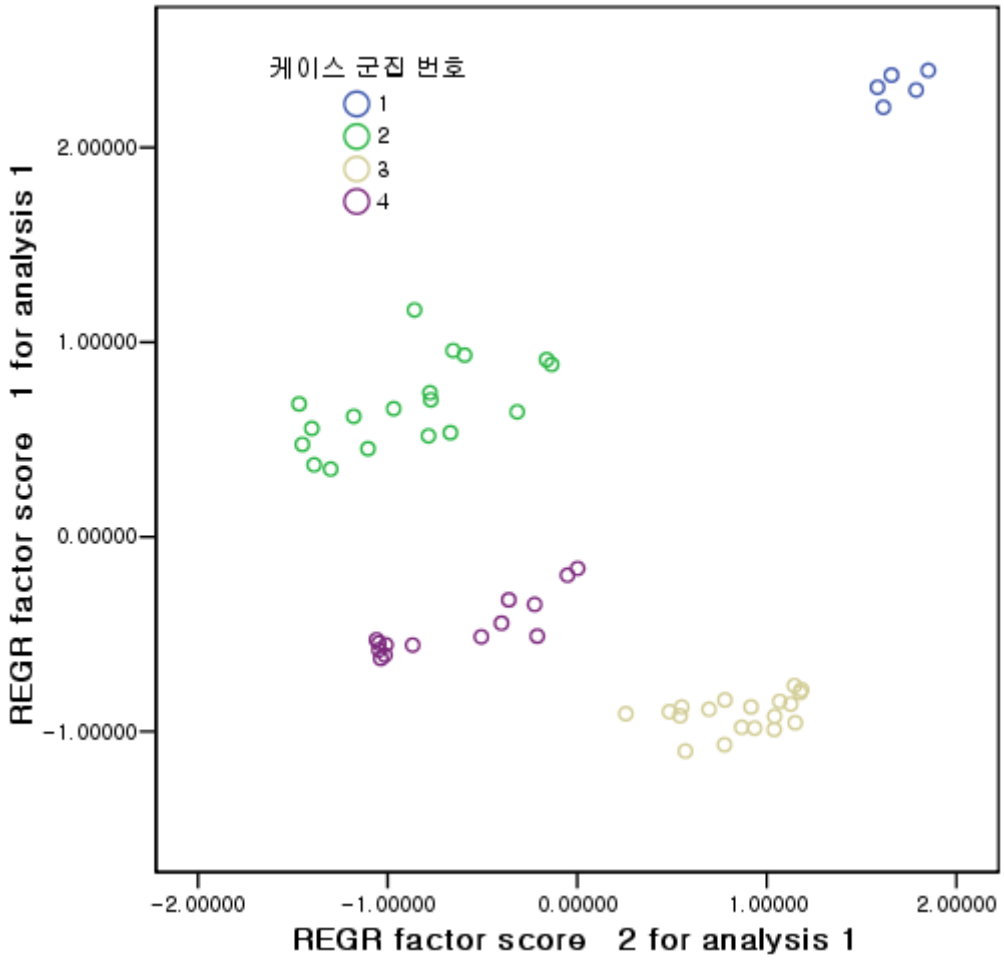
EXAMPLE 5-1 K-평균 집단 비계층적 군집분석

[Applied Multivariate Methods for Data Analysts, Dallas E. Johnson, 1998] ■PIZZA.SAV■



K-평균 군집분석에서는 다음 옵션만 설정하면 된다. 군집의 개수는 4 개로 설정하였다. 앞에서 구한 주성분에 의해 산점도를 그려보자. 계층적 군집분석과 결과가 유사해 보인다. 이는 계측적 군집분석에서 개체 분류 시 Centroid 방법을 사용했기 때문이다.


K-평균 군집분석
REGR factor score 1 1
REGR factor score 2 1
변수(V):
X1
X2
X3
X4
케이스 설명 기준변수(B):
ID
확인
명령문(P)
재설정(R)
취소
도움말
군집의 수(U):
2
방법
반복계산 및 분류하기(T)
분류만 하기(N)
K-평균 군집분석: 새 변수로 저장
소속군집(C)
군집중심으로부터의 거리(D)
계속
취소
도움말
저장(S)...
옵션(O)...


케이스 군집 번호
1
2
3
4
2.00000
1.00000
0.00000
-1.00000
REGR factor score 1 for analysis 1
-2.00000
-1.00000
0.00000
1.00000
2.00000
REGR factor score 2 for analysis 1

5.3 다차원척도법

다차원 척도법(MDS: Multi Dimensional Scaling)이란 n 개의 개체를 저 차원 가시적 공간(일반적으로 2 차원)에 나타내는 방법이다. 각 개체간 유사성(similarity) 혹은 거리는 저 차원으로 옮겨지더라도 원래 유사성 크기를 갖고 있어야 한다. 예를 들어 보자. Under-arm Deodorant 생산 회사에서 판매 전략을 세우기 위하여 각 제품들이 서로 얼마나 유사한지(가까운지) 알아보려고 한다. 이를 위하여 소비자를 임의로 선택하여 제품의 각 분야에 대한 평가를 하게 하였다. 즉 향기, 냄새 제거 정도, 사용 편리 정도, 옷에 묻어나는 정도 등을 10 점 만점으로 평가하였다. 이 점수들을 이용하여 제품의 유사성 정도를 2 차원 공간에 표현하는 방법이다.

개체간 유사성을 측정하는 방법은 metric 방법과 non-metric 방법이 있다.

(1)Euclidean distance ► 측정형 변수 거리 (Metric 방법)

(2)각 개체의 유사성(거리)을 사람들이 평가하도록 한다. (Metric/non-Metric 방법)

(3)평가자 들이 개체를 마음대로 분류하게 하고 빈도로부터 유사성을 측정한다. (non-Metric 방법)

응용 범위를 살펴보면 다음과 같다.

(1)회사들의 이미지 측정을 통한 고객 분류

(2)소비자들이 인지하고 있는 유사한 상품 속성이나 상품 분류에 사용

(3)인구 학적 특성, 경제적 특성을 기초하여 도시간 동질성 파악

5.3.1 유사성

다차원 척도법이란 n 개의 개체를 저 차원 가시적 공간(일반적으로 2 차원)에 나타낼 수 있도록 하는 방법이므로 각 개체간 거리(유사성)를 측정해야 한다. 군집분석과 유사해 보이지만 다차원 척도법은 개체의 유사성을 이차원에 표시하는 것이고 군집분석은 개체간의 거리(유사성)가 가까운 것끼리 묶어 가는 방법이다.

(1)metric 방법

metric 방법은 두 개체간의 거리(유사성)를 Euclidean distance 로 나타낸다.

측정변수 / 개체	X_1	X_2	⇢	X_p
1	x_{11}	x_{12}	⇢	x_{1p}
2	x_{21}	x_{22}	⇢	x_{2p}
⇣	⇣	⇣	⇢	⇣
n	x_{n1}	x_{n2}	⇢	x_{np}

두 개체 (D_1, D_2)간 거리는 $\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{ip}-x_{jp})^2}$

측정이 불가능한 경우는 각 개체간의 유사성(거리)을 리커드 척도(10 점, 100 점)를 이용하여 사람들이 평가하도록 하는 방법이 있다. 즉 각 x_1, x_2,…, x_p가 사용자들의 주관적인 평가 점수(리커드 척도)가 된다. Deodorant 제품을 분류할 경우 향기, 냄새 제거 정도, 사용 편리 정도, 옷에 묻어나는 정도 등을 10 점 만점으로 평가하였다면 이 점수가 개체들간의 거리를 측정하는데 사용된다.

(2)non-metric 방법

개체간의 거리를 사람들이 임의로 분류한 결과로부터 만들어진 빈도로 측정하는 방법이다. 이 방법을 이해하려면 예를 들어보는 것이 더 편리할 것이다. 50 명이 20 개의 개체를 임의로 분류하여 (1, 2)를 하나의 군집으로 분류한 사람이 30, (1, 20)을 하나의 군집으로 분류한 사람이 25 명, (2, 20)을 하나의 군집으로 분류한 사람이 45 명이라면

측정변수 / 개체	1	2	⇢	20
1	0			
2	30	0		
⇣	⇣	⇣	⇢	
20	25	45	⇢	0

이 경우 숫자가 클수록 거리는 가깝다. 즉 개체간 유사성이 높다. MDS 분석을 위하여 자료를 입력할 때는 (1-빈도/평가자수) 이것이 유사성이 된다.

5.3.2 기본 알고리즘

각 개체간 유사성을 측정한다. 개체의 개수가 n 개인 경우 k=n(n-1)/n 개 유사성 그룹이 존재한다.

(1)유사성이 작은 것부터 크기 순으로 배열한다. $S_{i1j1} < S_{i2j2} < \ldots < S_{ikjk}$

(2)개체를 m(일반적으로 2)차원으로 공간으로 줄일 경우 개체간의 거리를 구한다. 이는 물론 측정 변수 전체를 가지고 유사성을 측정한 2)와 다를 것이다. 2 차원 공간으로 줄일 수 있는지를 알아 보는 것이 STRESS 값이다.

$$STRESS = \sqrt{\frac{\sum_{i<j}\sum (S_{ij}^{2)} - S_{ij}^{3)})^2}{\sum_{i<j} (S_{ij}^{2)})^2}}$$

Stress	Goodness of fits
20%	Poor
10%	Fair
5%	Good
2.5%	Excellent

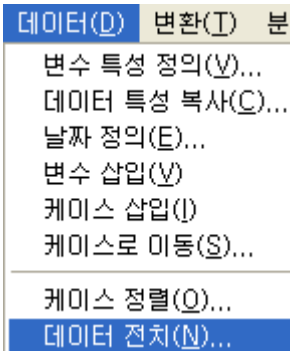
5.3.3 예제 1

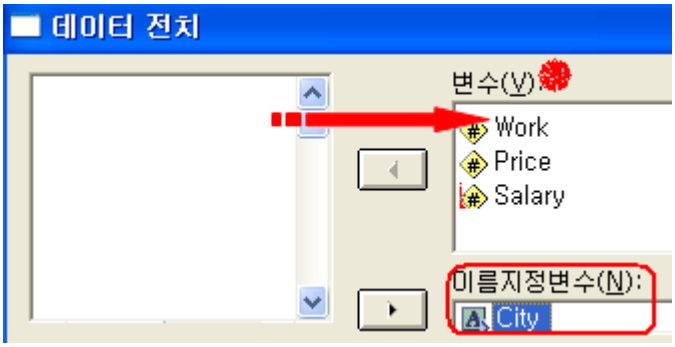
경제적 변인에 의해 도시를 군집화 하려 한다. 도시 이름, 12 개 직종 노동시간 가중평균, 물가, 시간당 임금 ■🖫CITY.SAV■

http://lib.stat.cmu.edu/DASL/Datafiles/Cities.html

City	Work	Price	Salary
Amsterdam	1714.00	65.60	49.00
Athens	1792.00	53.80	30.40
Bogota	2152.00	37.90	11.50
Bombay	2052.00	30.30	5.30

다차원 척도법에 의해 군집화 하려면 데이터를 전치(transpose)해야 한다. 이미 아래 상태로 입력되었다면 전치할 필요는 없지만...

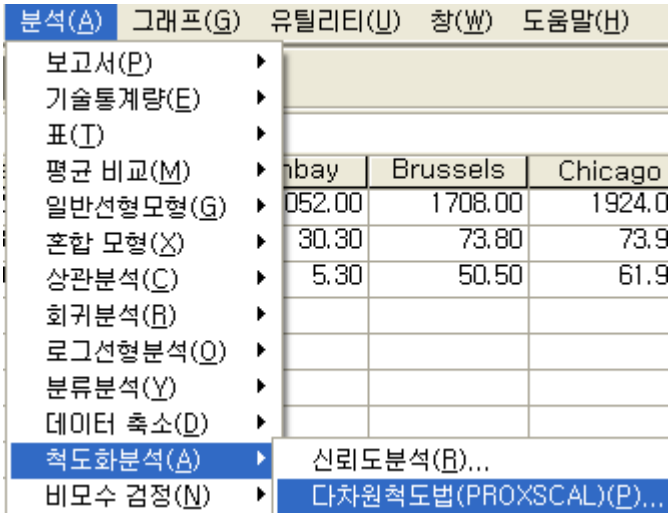


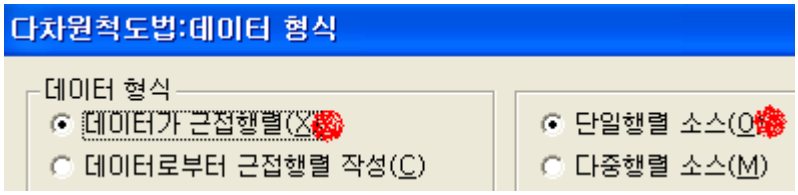


데이터가 전치되면 도시 이름이 변수이름이 되고 측정 변수 명은 CASE_LBL 이라는 변수명의 관측치로 저장된다.

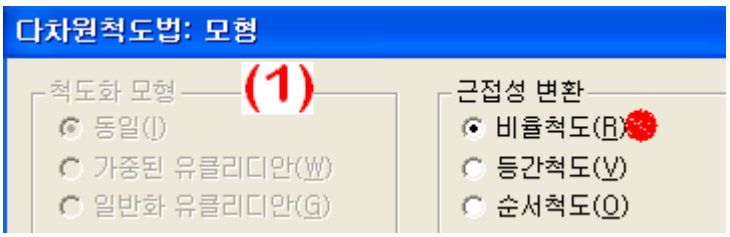
CASE_LBL	Amsterdam	Athens	Bogota	Bombay	Brussels	Chica
Work	1714.00	1792.00	2152.00	2052.00	1708.00	192
Price	65.60	53.80	37.90	30.30	73.80	7
Salary	49.00	30.40	11.50	5.30	50.50	6

이제 다차원 척도법에 의해 도시를 분류해 보자.





다른 옵션을 디폴트로 하고 모형에서만 측정 변수의 척도를 지정하면 된다. 예제 데이터의 경우 군집에 사용된 항목들이 측정형이므로 “비율 척도” 선택하면 된다.



스트레스 값이 0.001 이므로 군집이 매우 잘 되었다고 할 수 있다.

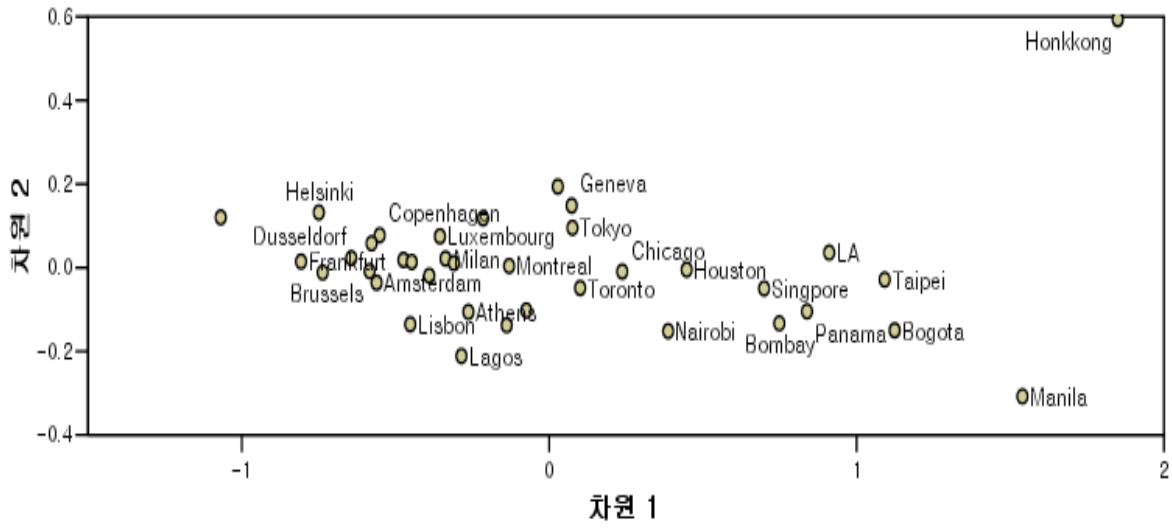
스트레스 및 적합도 측정

정규화된 원래 스트레스	,00083
스트레스-I	,02888[a]
스트레스-II	,04758[a]
S-스트레스	,00108[b]
설명된 산포	,99917
Turcker의 적합계수	,99958

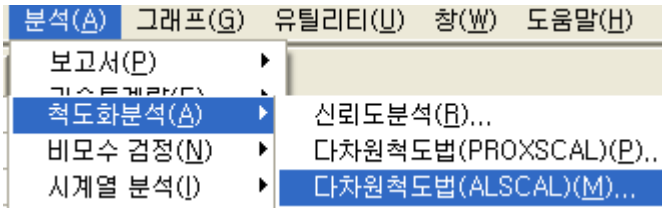
최종좌표

	차원	
	1	2
Amsterdam	-,561	-,036
Athens	-,263	-,106
Bogota	1,124	-,150
Bombay	,749	-,133
Brussels	-,584	-,008

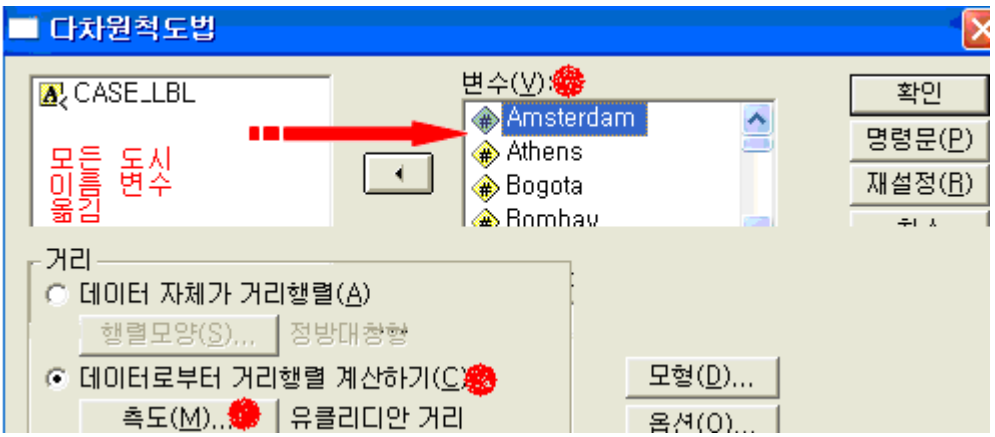
거리가 가까운 도시들은 경제 요인 ausdpyj 유사 도시라 할 수 있다. 마닐라나 홍콩은 다른 도시와 차별화 되고 있다. 왜? 이것은 수집된 데이터를 관찰하거나 변수간 산점도를 통하여 어떤 변수(항목)에서 다른 도시와 차이가 있는지 살펴보아야 한다.

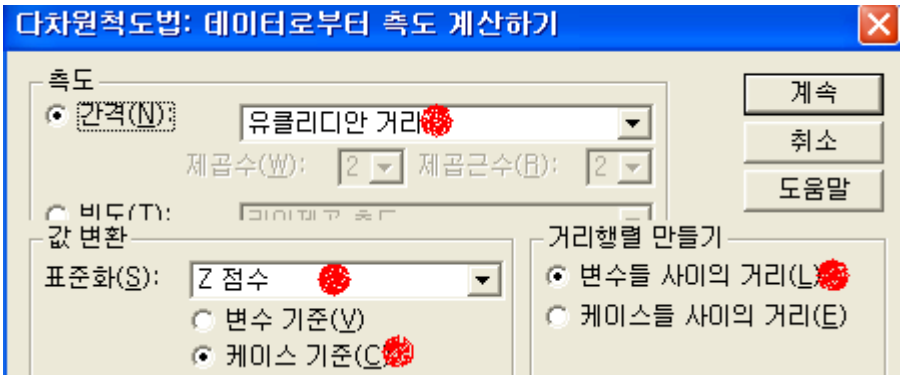


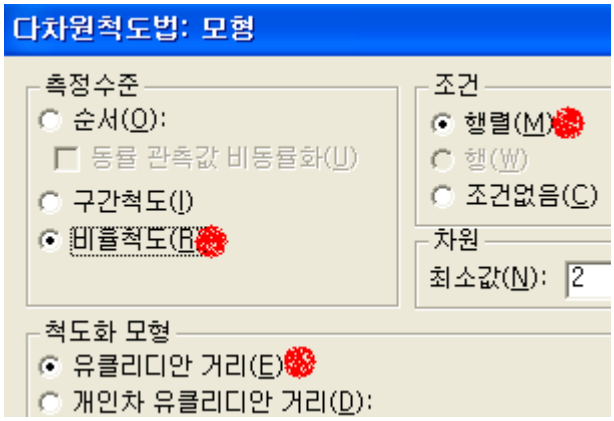
위의 예제의 경우 측정 단위가 다르므로 변수를 표준화 시켜 개체 간의 거리를 구하는 것이 바람직하다. 다차원척도법(PROXSCAL)(P)... 에는 표준화 설정이 없다. 측정 단위가 다를 때는 다음 다차원척도법(ALSCAL)(M)... 메뉴를 사용하자.

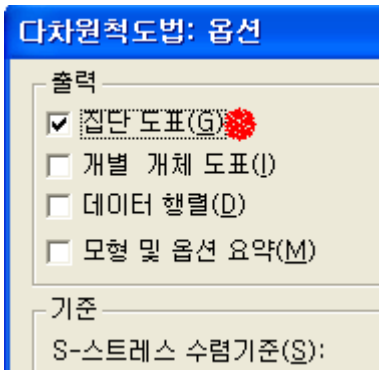


변수(개체)를 지정한다. 데이터로부터 거리 행렬을 계산하는 옵션을 선택한다. “측도” 옵션에서는 표준화 옵션을 선택하고 “케이스 기준”을 선택해야 한다. 옵션 메뉴에서 “집단 도표”를 선택하면 산점도가 그려진다.

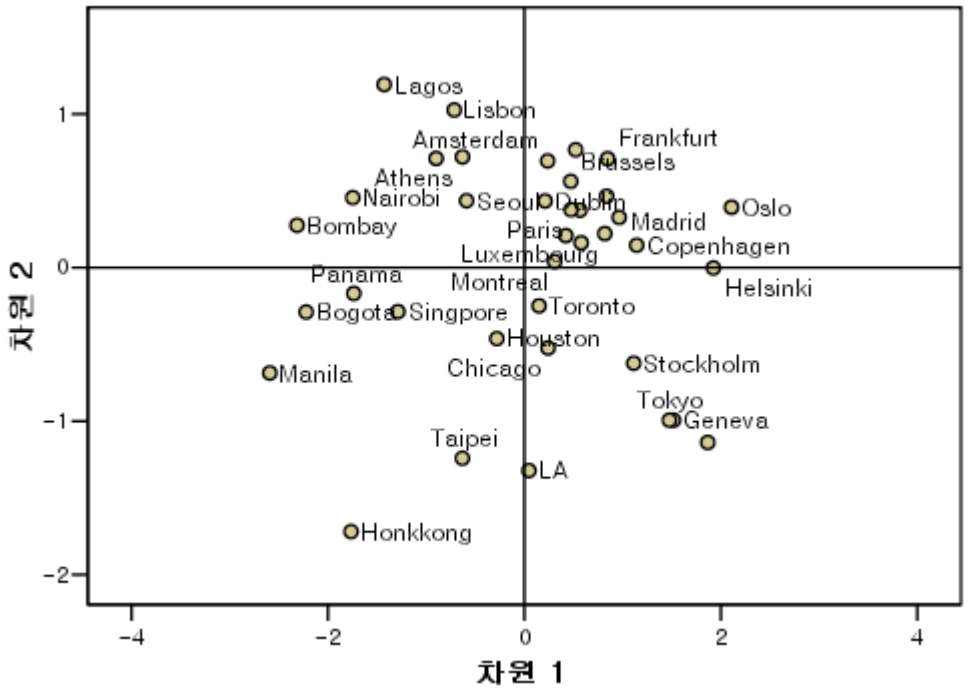








앞의 산점도와 다소 다르다. 이는 측정 변수(노동시간, 물가, 임금)가 표준화 되었기 때문이다. 여전히 홍콩은 다른 도시와 구별되고 LA 와 가장 가까운 도시는 Taipei 이다. 측정 단위가 다르다면 이 방법이 더 옳다.



5.3.4 예제 2

개체 간의 거리가 정방 행렬 형식으로 주어졌을 때 다음 절차에 의해 구하면 된다. 미국 도시 간의 거리를 입력한 자료이다. 이 자료는 다차원 척도법의 전형적인 예제 데이터이다. ■CITY1.SAV■

id	Atlanta	Chicago	Denver	Houston	LA	Miami	NY	SF	Seattle	DC
Atlanta	0	.	.	.	.	.	.	.	.	.
Chicago	567	0	.	.	.	.	.	.	.	.
Denver	1212	920	0	.	.	.	.	.	.	.
Houston	701	940	879	0	.	.	.	.	.	.
LA	1936	1745	831	1374	0	.	.	.	.	.
Miami	604	1188	1726	968	2339	0	.	.	.	.
NY	748	713	1631	1420	2451	1092	0	.	.	.
SF	2139	1858	949	1645	347	259	2571	0	.	.
Seatles	2182	1737	1021	1891	959	2734	2408	678	0	.
DC	543	597	1494	1220	2300	923	205	2442	2329	0

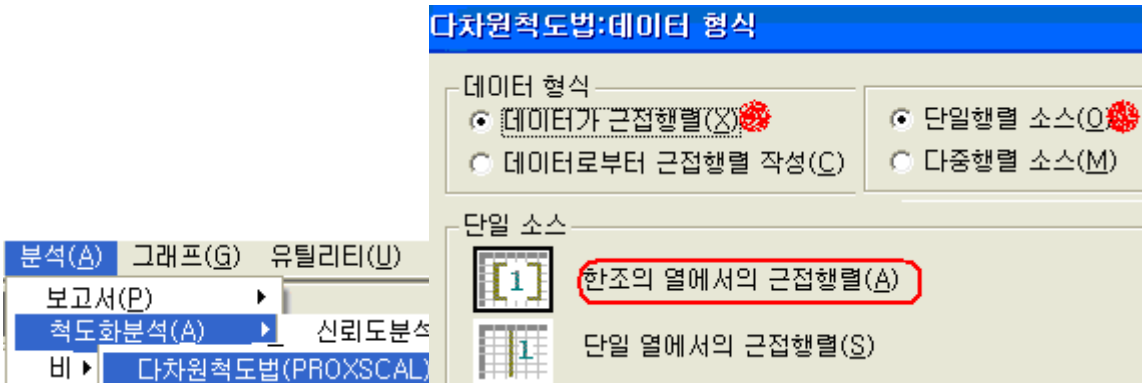
데이터 입력은 위와 같이 행렬의 형태로 입력한다. 첫 열의 데이터 이름과 2 열부터의 변수 이름은 일치해야 한다. 이 데이터는 “거리” 변수 하나만 측정한 형태이지만 여러 항목을 측정하는 경우에도 이런 형태의 데이터 입력이 빈번히 발생한다. 데이터 입력 방법은 자료 수집을 어떻게 하였느냐에 따라 달라진다.

①아이들에게 사탕, 아이스크림, 과자, 케이크, 과일이 대한 좋아함을 10 점 만점으로 측정하였다. ➔ 다차원척도법(PROXSCAL)(P)..., 등간 척도

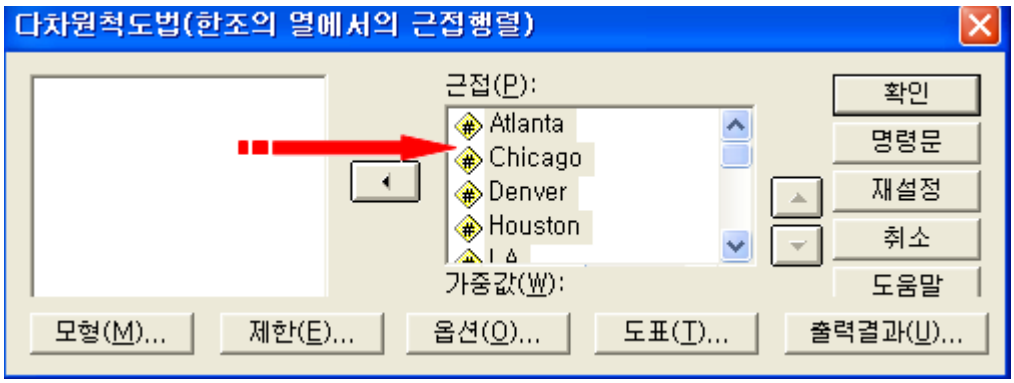
②아이들에게 사탕, 아이스크림, 과자, 케이크, 과일의 유사한 것부터 순위를 매겨라. 즉 사탕과 유사한 것 순위, 아이스크림과 유사한 순위,... 이렇게 하면 한 사람당 5X5 행렬이 생긴다. 그것을 사람 전체로 합하면 대칭 행렬이 된다. ➔ 다차원척도법(ALSCAL)(M)... 등간 척도

③아이들에게 사탕, 아이스크림, 과자, 케이크, 과일을 묶게 한다. 이렇게 하여 수집한 데이터로부터 (묶인 회수/전체응답자)를 대칭 행렬의 셀로 한다.

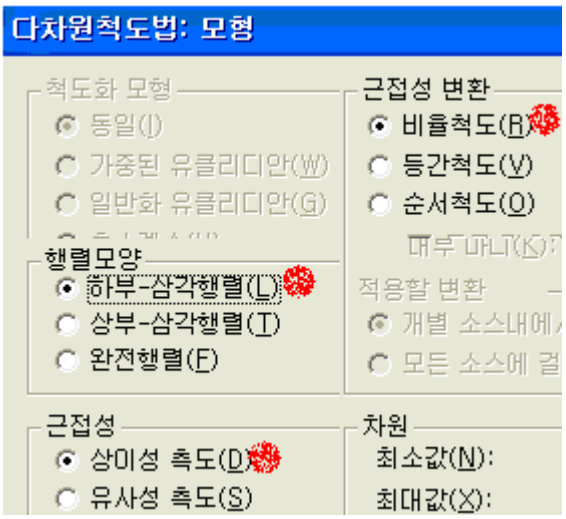
➔ 다차원척도법(ALSCAL)(M)... 등간 척도

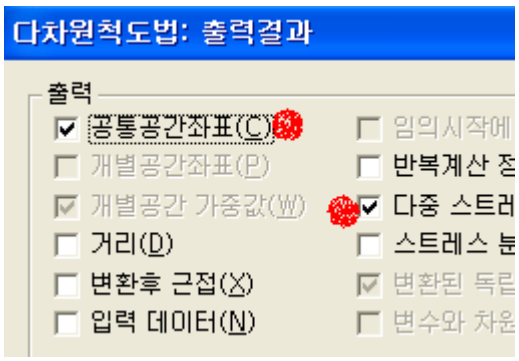


변수(개체 이름)을 "근접" 공간에 지정한다. 첫 열의 관측치와 변수 이름이 상이하면 다차원 척도 결과가 나타나지 않는다.



행렬 모양을 지정하고 근접성을 "상이성 측도"를 선택한다. 빈도의 경우는 "유사성 측도"로 하면 된다. 그런 경우 "비율 척도"는 선택할 수 없게 된다. 원하는 출력 옵션을 선택한다.





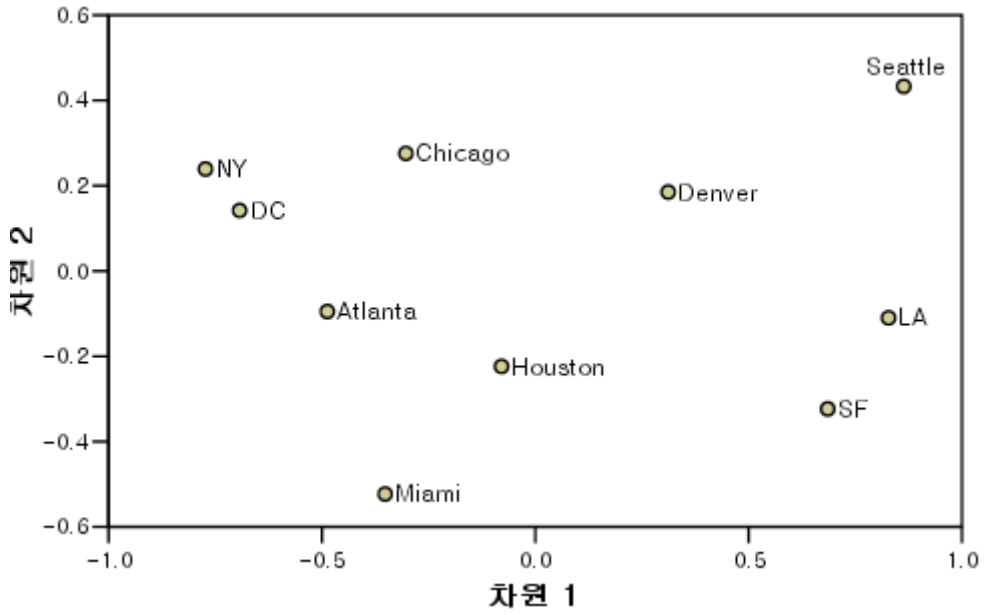
스트레스 값이 0.05 이므로 다차원 척도법에 의한 군집분석 결과는 믿을만하다. 가깝게 위치한 도시가 거리 개념에서 유사하다. 실제 지도 위치와는 전혀 다르다. Why?

스트레스 및 적합도 측정

정규화된 원래 스트레스	,03112
스트레스-I	,17640[a]
스트레스-II	,41449[a]
S-스트레스	,04926[b]
설명된 산포	,96888
Turcker의 적합계수	,98432

최종좌표

	차원	
	1	2
Atlanta	-,487	-,095
Chicago	-,303	,276
Denver	,311	,185
Houston	-,079	-,224
LA	,828	-,110



5.4 대응분석

다차원 척도법이 측정형 변수들에 의해 개체들의 유사성을 계산하고 이를 이용하여 개체들을 2 차원 공간에 나타낸 것이라면 대응분석(correspondence analysis)은 두 분류형 변수에 의해 만들어진 RXC 교차표(분할표)를 변수의 범주(수준)을 2 차원 공간에 표시하는 방법이다.

행(일반적으로 설명변수를 위치)의 프로파일을 $a_i = (\text{행퍼세트}1, \text{행퍼세트}2, ..., \text{행퍼세트}c)$ 이라 하자. $i = 1,2,...,r$ 각 행의 상대적 빈도(f_{+k}/n)를 계수로 한 주성분분석의 일종이다. 유사성의 척도인 거리는 다음과 같이 구한다.

$$\text{거리: } D = (a_i - a_j)' D^{-1} (a_i - a_j), \quad D = diag(\text{행1퍼센트}, ..., \text{행}q\text{퍼센트})$$

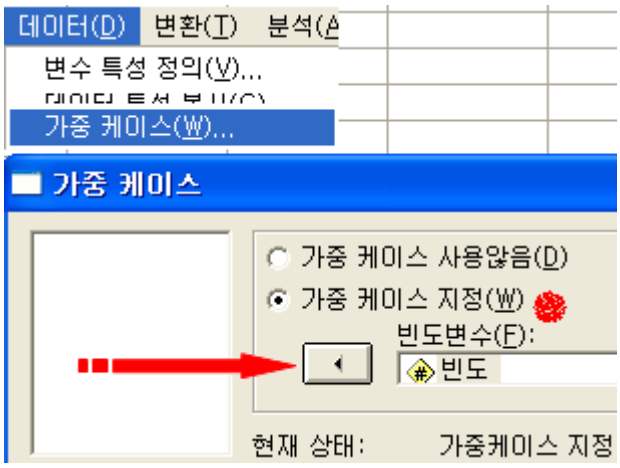
EXAMPLE 5-1 대응분석

직업별 정당(a, b, c, d) 선호도의 차이가 있는지 알아보기 위하여 조사한 결과 다음 데이터를 얻었다.응답자 총 수 n=1980 명

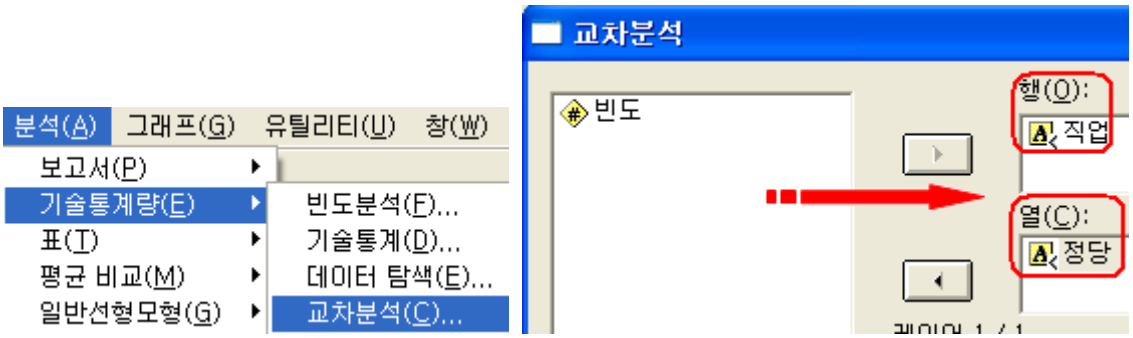
		정당			
		a	b	c	d
직업	기술직	185	240	335	130
	사무직	245	100	125	40
	전문직	105	60	70	25
	주부	40	35	75	40
	학생	40	25	30	35

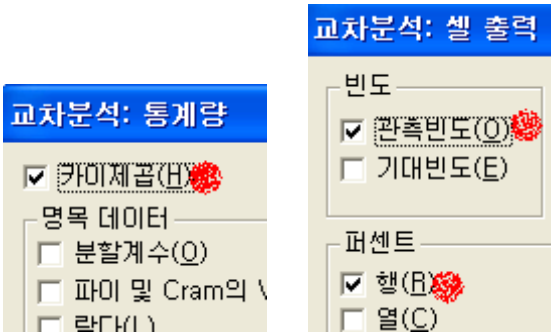
자료를 다음과 같이 입력하고 “빈도” 변수를 “가중치 케이스”로 지정한다. 그래야 데이터 개수가 1980 인 것으로 간주하여 분석한다.

직업	정당	빈도
학생	a	40
학생	b	25
학생	c	30
학생	d	35
주부	a	40
주부	b	35
주부	c	75
주부	d	40
사무	a	245



교차분석을 실시하자. 옵션은 다음과 같이 설정하면 된다.





피어슨 카이제곱 통계량의 유의확률 0.000 이므로 귀무가설(직업과 장당 선호도는 관계가 없다)이 기각되므로 직업별 정당 선호도의 차이는 있다고 결론 지을 수 있다. 어떤 차이? 이는 행 퍼센트에 의존한다. 기술직, 주부는 정당 C, 다른 직종은 정당 A 를 가장 선호한다. 다소 주관적인 여러 방향의 해석이 가능하다.

직업 * 정당 교차표

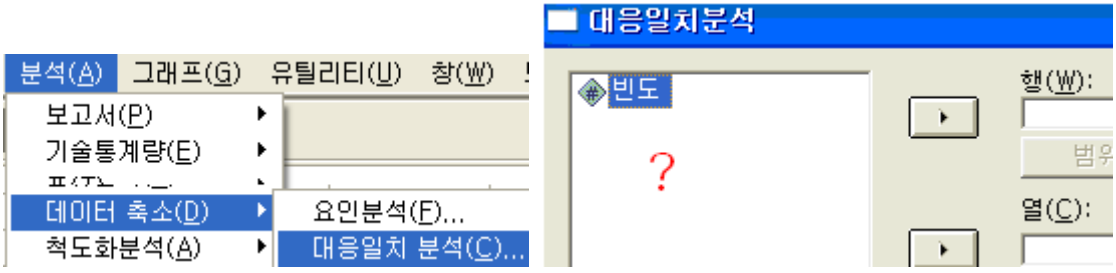
			정당				전체
			a	b	c	d	
직업	기술직	빈도	185	240	335	130	890
		직업의 %	20.8%	27.0%	37.6%	14.6%	100.0%
	사무직	빈도	245	100	125	40	510
		직업의 %	48.0%	19.6%	24.5%	7.8%	100.0%
	전문직	빈도	105	60	70	25	260
		직업의 %	40.4%	23.1%	26.9%	9.6%	100.0%
	주부	빈도	40	35	75	40	190
		직업의 %	21.1%	18.4%	39.5%	21.1%	100.0%
	학생	빈도	40	25	30	35	130
		직업의 %	30.8%	19.2%	23.1%	26.9%	100.0%
전체		빈도	615	460	635	270	1980
		직업의 %	31.1%	23.2%	32.1%	13.6%	100.0%

카이제곱 검정

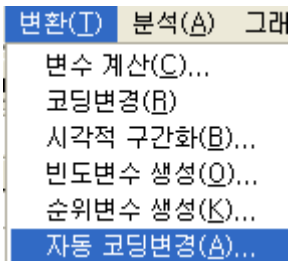
	값	자유도	점근 유의확률 (양측검정)
Pearson 카이제곱	169.120[a]	12	.000
우도비	164.952	12	.000
유효 케이스 수	1980		

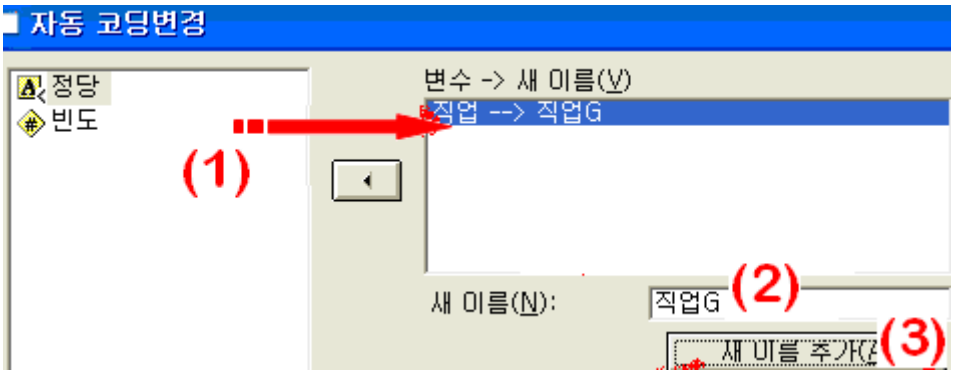
a. 0 셀 (.0%)은(는) 5보다 작은 기대 빈도를 가지는 셀입

행과 열의 변수 범주의 군집화를 위하여 대응 분석을 실시해 보자. 대응분석을 하려면 행과 열 변수도 숫자로 입력되어 있어야 한다. 그냥 해보자. 변수 목록에 행과 열 변수 명이 나타나지 않는다. 열과 행 변수를 숫자형 변수로 변환하여야 한다.



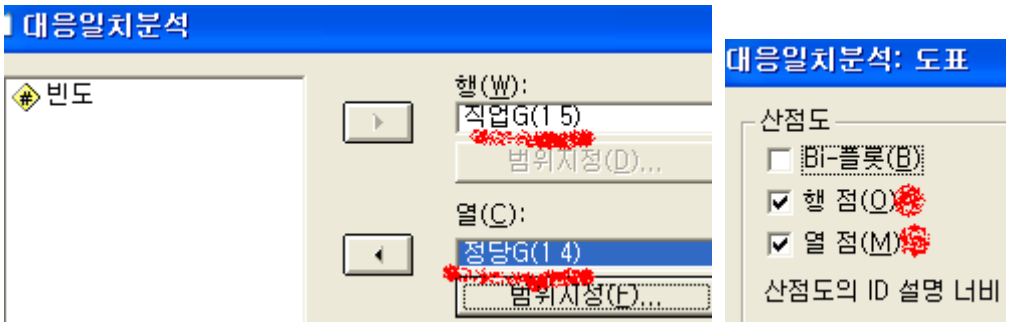
행 변수 변환 절차는 다음과 같다. (1)변환하려는 변수를 지정하고 (2)새 이름을 적어 (3) 버튼을 누른다. 열 변수도 동일하게 하면 된다. 코딩 변경이 끝난 후 "변수 보기"(시트 왼쪽 아래) 폴더를 열면 코딩 변경 값이 저장된 것이 보인다.



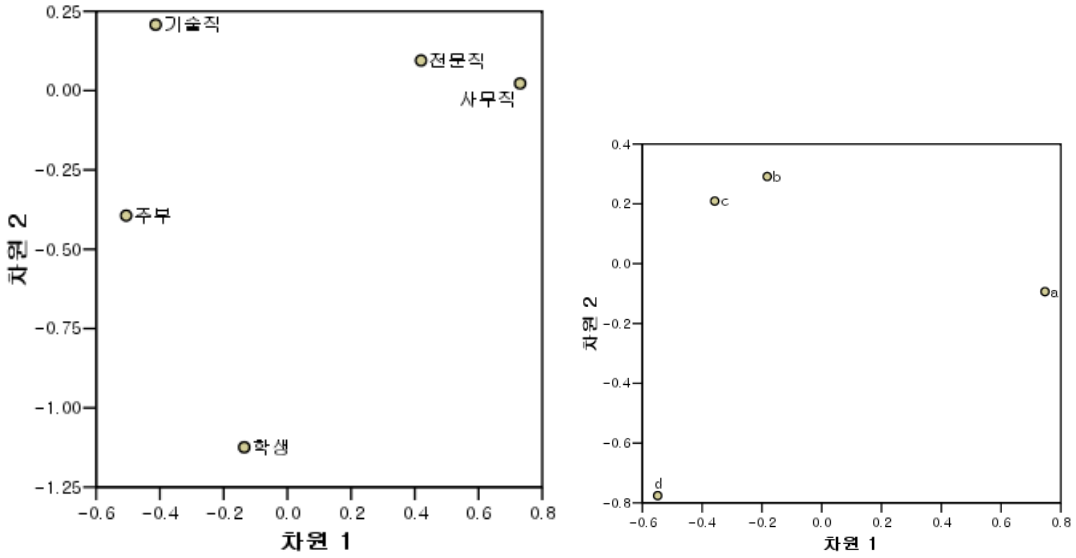


이름	유형	자리수	소수점이하	설	값
직업	문자열	8	0		없음
정당	문자열	8	0		없음
빈도	숫자	4	0		없음
직업G	숫자	1	0		{1, 기술직}...
정당G	숫자	1	0		{1, a}...

코딩 변경이 끝나면 행과 열에 변수를 지정하고 "범위지정"도 한다. 다른 옵션은 모두 디폴트로 하고 "도표" 창을 열어 아래와 같이 지정하고 실행한다.



정당 선호도 면에서 보면 전문직과 사무직은 유사 개체이다. 직업 측면에서 보면 B, C 가 유사 정당이다.



EXERCISE

Fisher 의 Iris(꽃) 데이터이다. 이것에 대해 4 가지 특성을 조사하였다. 물론 이 꽃은 어떤 종인지 (VARIETY: S, C, V ► 3 종류) 이미 알고 있지만 종을 모른다고 가정하고 군집분석(계층적방법, 비계층적 방법)을 실시해 보자. 꽃잎 길이(petal length), 꽃잎 넓이(petal width), 수술 길이(stamen length), 수술 넓이(stamen width) ■🖫IRIS.SAV■

아이스크림이 31 개 종이 있다. 이들을 다차원 척도법에 의해 군집화 하려고 한다. 자료수집 방법과 데이터 입력 방법을 논의해 보자.