Multivariate Data Analysis (SPSS 중심)

통계교육원

2009.05.20-21

한남대학교 권세혁 교수
042.629.7622 http://wolfpack.hnu.ac.kr

■ 다변량(Multivariate Data Analysis) 개념
- 변수가 2개 이상인 데이터 분석 방법 (넓은 의미)

■ 인과 관계 분석 (넓은 의미) $Y=f(X_1, X_2, ..., X_p)+e$

■ 데이터 차원 축약: 변수들의 상관관계 활용
- 주성분분석 (principal component analysis)
- 요인분석 (factor analysis)

■ 개체 분류: 개체 유사성(거리) 이용
- 군집분석 (clustering analysis)
- 판별분석 (discriminant analysis)

■ Other MDA
- 정준상관분석 (canonical correlation): 변수 그룹간 상관분석
- 대응분석 (correspondence analysis): 개체 분류 범주 분류

←측정변수→

↑ 개체 ↓

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ & \cdots\cdots\cdots & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

개체 분류
- 군집 분석
- 판별 분석

데이터 축약
- 주성분 분석
- 요인분석

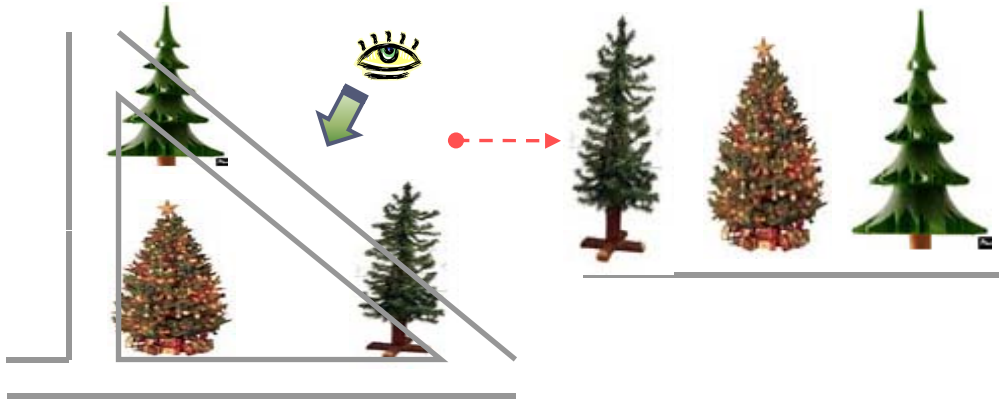
	주성분 분석	요인 분석	판별 분석	군집 분석
변수들 관계 탐색	S	D	N	N
자료 탐색	D	S	N	S
새 변수 만들기	Yes	Yes	No	No
개체 분류	No	No	Yes	Yes
그룹간 평균 비교	P	P	R	R
변수 그룹	P	P	N	N
차원(변수) 줄이기	D	P	N	N

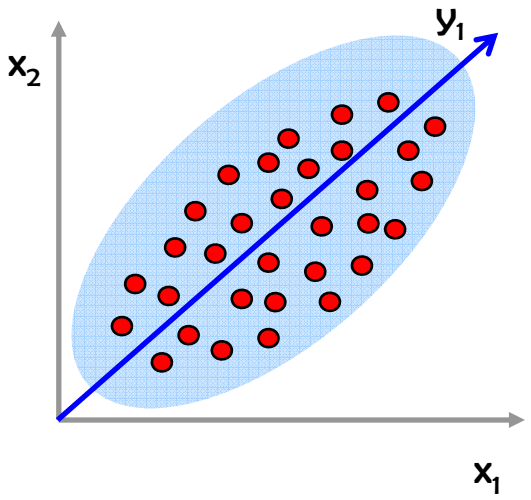
- Definitely, Sometimes, Rare, Possible, Not possible
- Non-metric 군집분석: MDS

주성분분석	요인분석	판별분석	군집분석
다변량 변수의 상관관계를 이용하여 데이터 차원 축소	소수 몇 개의 공통 인자를 이용하여 다변량 데이터 변수 분류	소속 집단이 알려진 개체의 측정 다변량변수 이용하여 판별식을 얻고 새로운 개체 분류	측정 다변량변수를 활용하여 개체들의 유사성을 계산하고 이를 이용하여 개체를 분류

개념

- 주성분 분석은 데이터가 흩어져 보이는 방향을 찾아서 그로부터 데이터를 분석하는 기법
- 데이터가 가진 정보(변동)을 잘 간파할 수 있는 방향에서 데이터를 재표현
- 데이터의 분산을 최대로 하는 합성변량을 찾음으로써 파악 가능함
- 변수들이 가진 변동(정보)을 일부 희생하여 변수를 축약하는 방법
- (예제 그림) 2차원 데이터 정보(변동)를 1차원에 표현? 어느 방향에서 데이터를 바라보나?





- 데이터의 변동을 크게 바라보는 시각: 주성분
- 차원 축소: p개 원변수 변동을 1~2개의 주성분 변동으로 설명

■예제 데이터 (p=2, 개체=20)

■원변수 산점도

- 개체에 대한 측정변수의 변동(정보) 표현
- 공분산

	IQ	math
IQ	18.6184	
math	56.1974	459.1026

■주성분 Y=LX, L=loading matrix

- $Y1=l_{11}*IQ+l_{12}*Math$
- $Y2=l_{21}*IQ+l_{22}*Math$

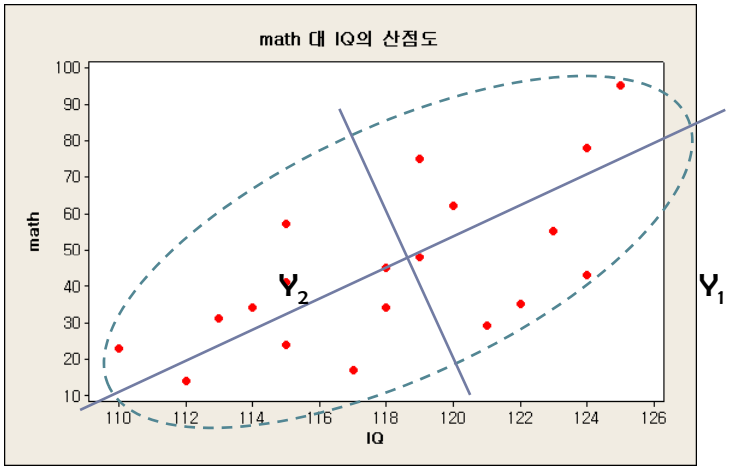
L1	L2
0.12	0.99
0.99	-0.12

■주성분 산점도 및 공분산

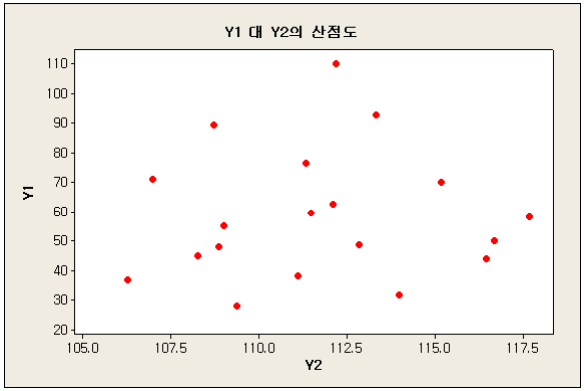
공분산: Y1, Y2

	Y1	Y2
Y1	466.1593	
Y2	0.0000	11.5618

Y1와(과) Y2의 Pearson 상관 계수 = 0.000
P-값 = 1.000



IQ와(과) math의 Pearson 상관 계수 = 0.608
P-값 = 0.004



IQ	Math
110	23
115	24
125	95
112	14
118	34
124	78
121	29
120	62
115	57
117	17
119	48
123	55
118	45
121	29
113	31
115	41
119	75
114	34
122	35
124	43

Typical example

- 기성복 바지 구입: 여러 측정치들이 (길이, 둘레) 2개의 이미지 변수로 축약
  - 이미지 변수는 합성변수, 주성분변수라 한다.

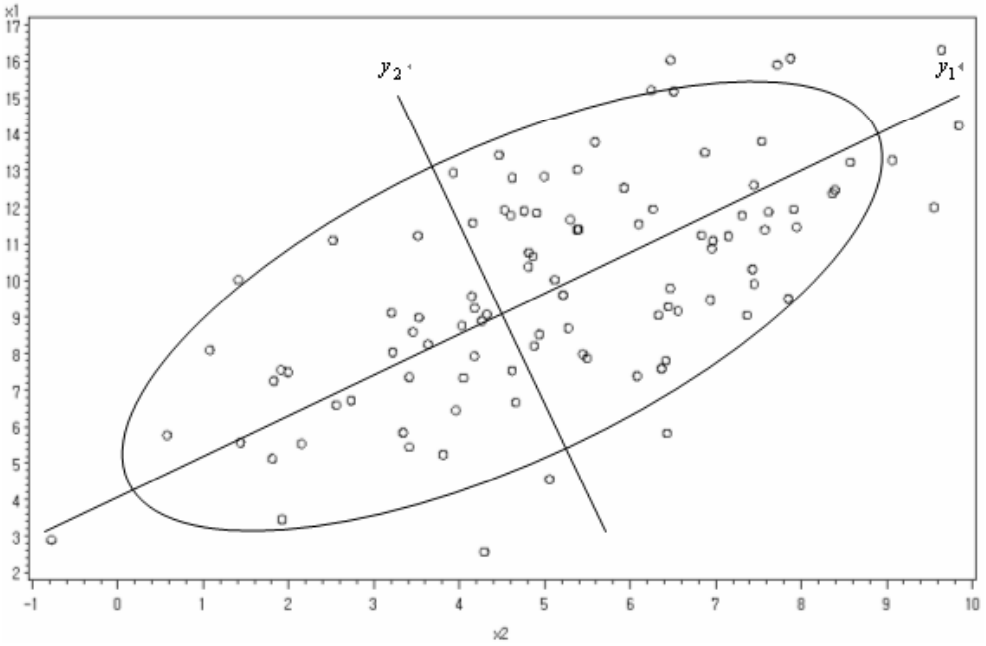
활용 방법: (p≥3) 다변량 데이터를 저차원(1, 2 차원)으로 축약한 후

- 주성분 분석은 중간 단계이다. (데이터 축약을 통한 개체 분류나 변수 구조 탐색 도구)
- 데이터 스크린 (*)
  - 상자 수염그림이나 산점도를 이용하여 개체 특성 표현 가능
- 개체 분류 (*)
  - 합성변량(이를 주성분)에 의해 개체 분류 가능
- 개체 순위 (*)
  - 다변량 측정변수 개념 하에서 개체 순위부여 가능: 가중 평균 개념
- 개체 판별
  - 판별 변수가 많은 경우 변수를 축약하여 이를 이용: 컴퓨팅 기술 발전 후 사용하지 않음
- 회귀분석
  - 다중공선성 문제 해결: 주성분 변수는 서로 독립이라는 성질 이용
  - not recommended: 주성분 변수 성격이 모호

■원 변수((x1, x2) 이변량 정규분포) 공분산 행렬

$$\Sigma = \begin{bmatrix} 9 & 2 \\ 2 & 4 \end{bmatrix}$$

- 상관계수 0.33



■합성변량(이를 주성분이라 한다)

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{bmatrix} 0.94 & 0.33 \\ -.33 & 0.94 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

■주성분 성질

- 주성분 변수의 분산의 합 (9.7, 3.2)
  - 원변수 분산의 합과 동일하다.
- 주성분 변수 상관계수 0
- 주성분 변수는 원변수의 선형결합
- 주성분은 데이터 변동을 가장 잘 표현하는 순서대로 구해진다.
- 그러므로 최초 1-2개 (최대 3개) 활용하여 원변수를 축약한다. (예제: Y1으로 (X1, X2) 정보 표현)

주성분(principal component)이란?

- 주성분은 원변수의 선형결합이다. (합성변량) Y=LX
- 산형계수 행렬 L: 부하행렬(loading matrix)

$$\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} l_{11}\, l_{12} \cdots l_{1p} \\ l_{21}\, l_{12} \cdots l_{2p} \\ \vdots \\ l_{p1}\, l_{p2} \cdots l_{pp} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = L\underline{x}$$

주성분 개념

- p개 원 변수의 선형 결합인 주성분 변수를 이용하여 원 변수의 공분산 구조를 설명하는 방법
- 공분산 구조를 설명한다는 것은 원 변수의 변동 합과 주성분 변수의 변동 합은 동일하다는 것을 의미한다.

계산 원칙

- 주성분 변수는 서로 독립이다.
- 주성분의 개수는 원변수의 개수 p와 동일하다.
- 첫 번째 주성분의 분산(변동, 원변수 변동에 대한 설명 능력)은 가장 크고, 순차적으로 줄어든다.
  - 제일 주성분 (제이 주성분 포함)으로 원변수 변동의 대부분을 설명: 80% rule

■원변수 벡터와 공분산 행렬 (covariance matrix, 모집단 Σ, 표본 S)

$$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots \sigma_{2p} \\ \cdots & & \\ \sigma_{p1} & \sigma_{p2} & \cdots \sigma_{pp} \end{bmatrix}$$

■공분산 행렬(Σ)의 고유치

- 계산 방법: $|\Sigma_{p\times p} - \lambda I_p| = \underline{0}$ 을 만족하는 λ들을 고유치라 한다. $\lambda_1 \geq \lambda_2 . \geq .. \geq \lambda_p$
- 고유치는 행렬의 차수만큼 존재한다.
- 고유치 의미: 주성분 변수의 분산 (슬라이드 9에서 타원의 폭과 높이에 해당)

■공분산 행렬(Σ)의 고유벡터

- 계산방법: $\Sigma \underline{e}_i = \lambda_i \underline{e}_i$ 을 만족하는 벡터 e를 고유벡터라 한다.
- 고유벡터는 무수히 많이 존재
- 고유벡터는 서로 orthogonal 하다. $\underline{e}_i' \underline{e}_j = 0 \ \ for\ i \neq j$
- 고유벡터를 주성분 계산 선형계수로 사용한다. (e=l)

$$L = \begin{pmatrix} l_{11}\, l_{12} \cdots l_{1p} \\ l_{21}\, l_{12} \cdots l_{2p} \\ \vdots \\ l_{p1}\, l_{p2} \cdots l_{pp} \end{pmatrix} = \begin{pmatrix} e_{11}\, e_{12} \cdots e_{1p} \\ e_{21}\, e_{12} \cdots e_{2p} \\ \vdots \\ e_{p1}\, e_{p2} \cdots e_{pp} \end{pmatrix}$$

주성분 구하기

- 제일 주성분 (first principal component)
  - $\underline{a}_1'\underline{a}_1 = 1$을 만족하는 벡터 중 $V(\underline{a}_1'(\underline{x}-\underline{\mu}))$을 최대화 하는 벡터 $\underline{a}_1$ 를 선형계수로 하여 구해진 합성변수. $\underline{y}_1 = \underline{a}_1'(\underline{x}-\underline{\mu})$
  - 슬라이드 12: 공분산 행렬(S)로부터 얻어진 고유치 λ1에 대응하는 고유벡터 e1 중 $\underline{e}_1'\underline{e}_1 = 1$을 만족하는 고유벡터
- 제이 주성분 (second PC) $\underline{y}_2 = \underline{a}_2'(\underline{x}-\underline{\mu})$
  - $\underline{a}_1'\underline{a}_2 = 0, \underline{a}_2'\underline{a}_2 = 1$을 만족하는 벡터 중 $V(\underline{a}_2'(\underline{x}-\underline{\mu}))$ 을 최대화 하는 벡터 $\underline{a}_2$ 를 선형계수로 하여 계산한 합성변수.
  - 슬라이드 12: 공분산 행렬(S)로부터 얻어진 고유치 λ2에 대응하는 고유벡터 e2 중 $\underline{e}_1'\underline{e}_2 = 0, \underline{e}_2'\underline{e}_2 = 1$ 을 만족하는 고유벡터
- 주성분 개수는 원변수 개수와 동일하다.

주성분 성질2

$$Var(Y_i) = \underline{e}_i'\Sigma\underline{e}_i = \lambda_i \qquad Cov(Y_i, Y_k) = \underline{e}_i'\Sigma\underline{e}_k = 0, \ for\ i \neq k$$

$$\sum_{i=1}^{p} V(Y_i) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} V(X_i)$$

- 주성분의 변동은 고유치와 같고, 변동의 크기는 제일, 제이, ... 순이다.
- 주성분 변수는 사로 독립이다.
- 주성분 변수의 변동 합은 원변수 변동 합과 동일하다.
- 주성분 변수는 계산 식을 실제 데이터에 의해 계산된 값을 주성분 점수(score)라 한다.

주성분 기여율

- 원변수 변동의 설명능력 측정: (주성분 변수의 분산)/(원변수 변동 합)
- K번째 주성분 변수 기여율

$$\frac{S_{y_1}^2}{S_{x_1}^2 + S_{x_2}^2 + \ldots + S_{x_p}^2} = \frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}$$

누적 기여율

- 제일 주성분부터 K번째 주성분까지 변동 합

$$\text{누적기여율} = \sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$$

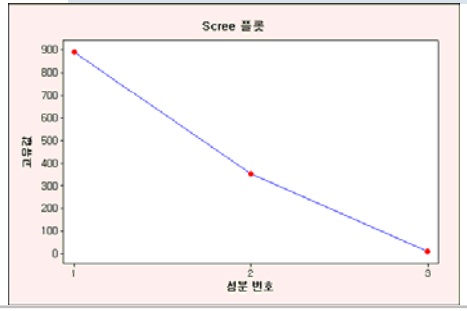
상관계수 행렬 (correlation matrix, R) 사용

- 원변수의 측정 단위가 상이한 경우: 주성분 계산 시 단위 크기가 큰 원변수의 영향(분산이 크므로)이 크다.
- 문제 해결을 위하여 상관계수 행렬 사용하여 고유치, 고유벡터를 구한다. (절차는 공분산 행렬과 동일)

$$\text{누적기여율} = \sum_{i=1}^{k} \lambda_i / p$$

주성분 개수 결정

- 80% rule (공분산 행렬 사용 시) / 고유치 1 이상 (상관계수 행렬 사용 시)
- Scree 도표 이용
  - Y축 고유치, x 축을 주성분 순차 번호 산점도
  - 고유치가 감소 경향을 시각적으로 표현
  - 급격히 감소하는 곳에서 주성분 개수 결정



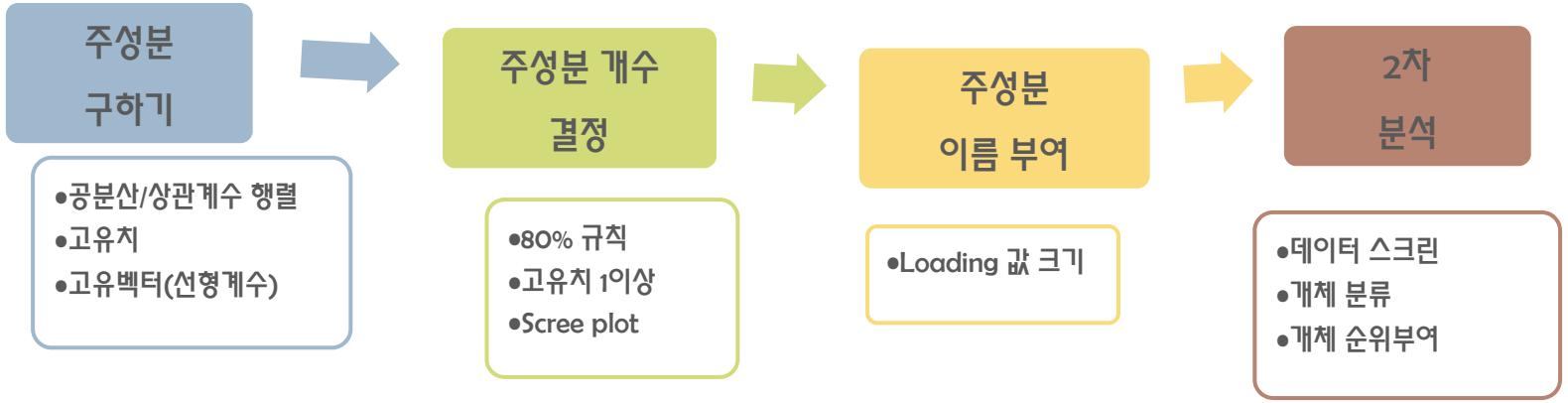
■부하 (loading) 정의

- 공분산 행렬로부터 얻어지는 $\underline{c}_j = \sqrt{\lambda_j}\,\underline{e}_j$ 을 성분부하(component loading)라 정의
- 주성분 변수를 계산할 때 사용되는 선형계수 값

■사용

- 주성분 변수를 계산할 때 원변수가 주성분 변수에 미치는 영향 정도가 부하이다.
- 부하 값이 크다는 것은 원변수의 영향력이 크다. (*전제조건: 원변수의 단위가 유사, 그렇지 않으면 상관계수 행렬 사용)
- 이를 이용하여 주성분 변수의 이름 부여한다.

■주성분 분석 절차

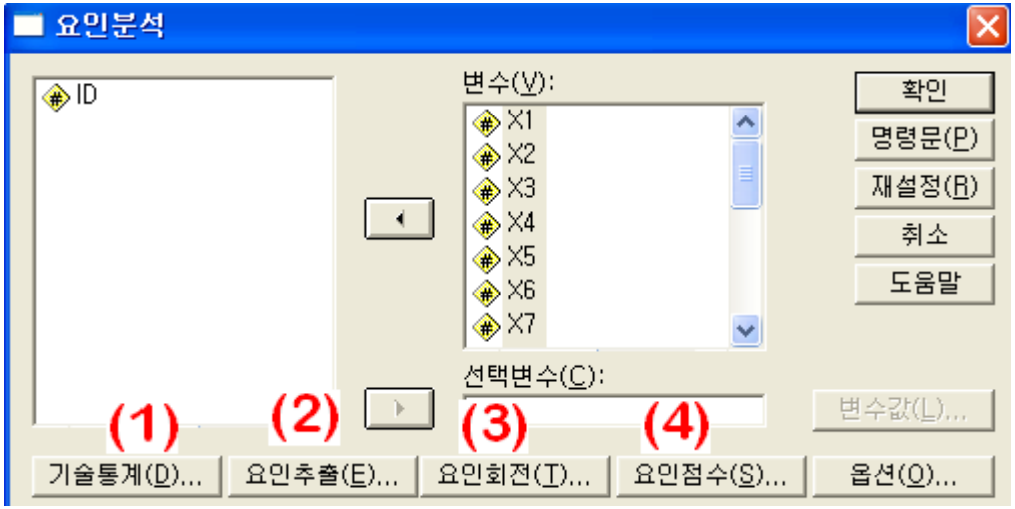


■데이터 APPLICANT.SAV

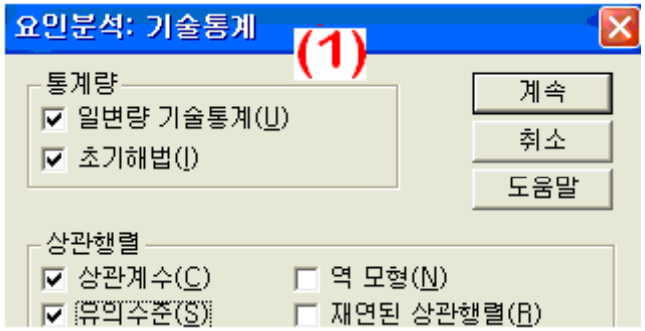
- A 회사에서는 48명의 지원자에 대해 그들의 능력을 10점 만점으로 15개 영역(변수 이름 X1~X15)을에 대해 조사

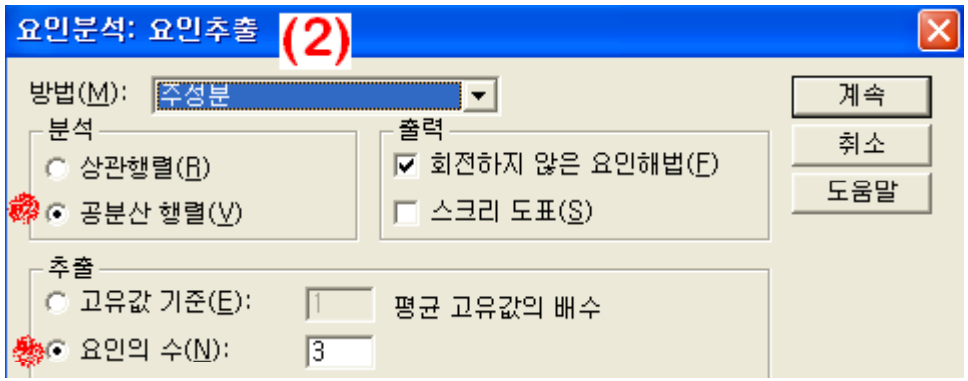
■메뉴

- 주성분 분석은 요인분석의 일부분으로 되어 있음



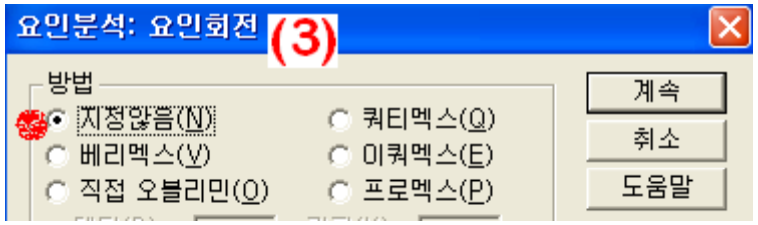
■메뉴 (cont.)

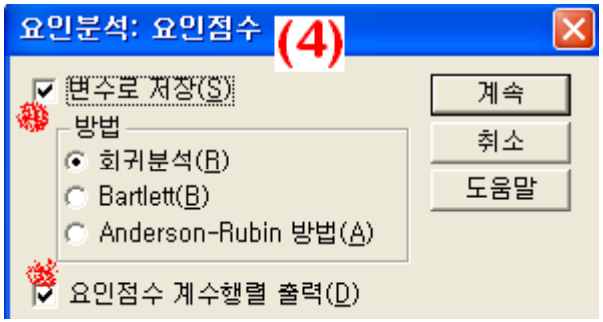




- 변수의 단위나 속성이 다르면 상관행렬, 동일하면 공분산 행렬 사용
- 주성분 개수는 지정하거나 고유값=1 기준(80% 규칙과 일치) 선택하면 된다.
  - 사후적으로 선택할 수 있음. 우선 3개로 초기 지정

■메뉴 (cont.)





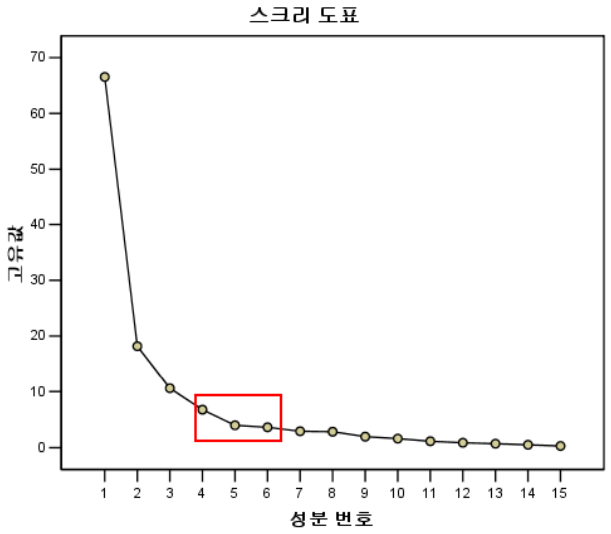
- 요인 회전
  - 고유치에 대응하는 고유벡터는 무수히 많이 존재한다는 사실 활용
  - 부하 값의 구별 용이하게 하기 위하여

- 계수행렬=부하행렬
  - 주성분 변수(점수) 이름 부여 시 사용

■변동 설명 비율 및 Scree 도표

설명된 총분산

성분	초기 고유값[a] 전체	% 분산	% 누적
1	66,536	54,300	54,300
2	18,181	14,837	69,137
3	10,591	8,643	77,780
4	6,768	5,523	83,303
5	3,986	3,253	86,556
6	3,628	2,961	89,517
7	2,916	2,380	91,896
8	2,836	2,314	94,210
9	1,955	1,596	95,806
10	1,613	1,317	97,123
11	1,136	,927	98,050
12	,873	,712	98,762
13	,707	,577	99,339
14	,509	,415	99,754
15	,301	,246	100,000



- 주성분은 원 변수의 개수(15)만큼 존재한다. "전체" 열은 고유치의 값이고, "%분산" 열은 주성분의 원 변수 변동 설명 비율이다. 마지막 "%누적" 열은 주성분 설명력의 누적이다. 출력 결과에 의하면 4개의 주성분이면 누적 설명력이 83%이므로 15개의 변수가 4개의 주성분으로 축약될 수 있다. 그러나 주성분이 3개인 경우 설명력은 78%으로 80%와 차이가 없고 주성분 개수가 적을수록 주성분분석 효율성이 높아지므로 3으로 하는 것이 적절하다.
- 상관계수 행렬 이용 시 고유치 1 이상과 80% 규칙은 일치

부하 값

성분점수 계수행렬[a]

	성분		
	1	2	3
X1	,049	,233	,165
X2	,032	-,014	,025
X3	,007	,047	-,080
X4	,070	-,061	,534
X5	,069	-,134	-,141
X6	,131	-,146	-,121
X7	,037	-,179	,348
X8	,160	-,073	-,298
X9	,067	,494	,025
X10	,114	,009	-,103
X11	,112	-,084	-,221
X12	,126	-,053	-,047
X13	,139	-,019	,040
X14	,074	-,028	,315
X15	,111	,364	,017

제일주성분 $Y_1 = 0.049 * X1 + 0.032 * X2 + ... + 0.111 * X15$

제이주성분 $Y_2 = 0.233 * X1 - 0.014 * X2 + ... + 0.364 * X15$

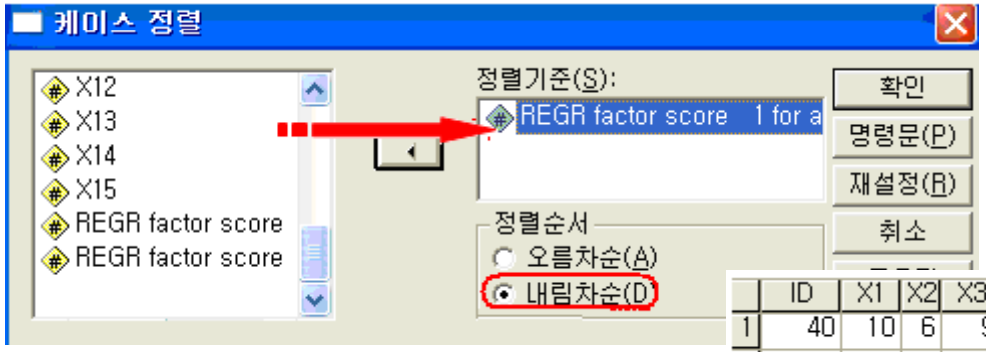
- 주성분 이름 부여 (not easy)
  - 제일 주성분 계수 크기에 의하면 X6(명석), X8(판매능력), X10(돌파력), X11(야망), X12(개념파악), X13(잠재력) 변수의 크기가 크므로 제 1 주성분은 정신적&지적 능력으로 할 수 있다.
  - 제이 주성분에서는 X9(경험), X15(적합)이 큰 역할을 하므로 경험 주성분이라 할 수 있다.
  - 제삼 주성분에서는 X4(호감), X7(진실), X14(사교)의 크기가 크므로 감성 변수로 이름하면 된다.

■주성분 점수 활용

▪ 15개의 원 변수를 축약하여 3개의 주성분을 얻었다. 데이터에 "FAC*_*" 라는 이름으로 저장된다.

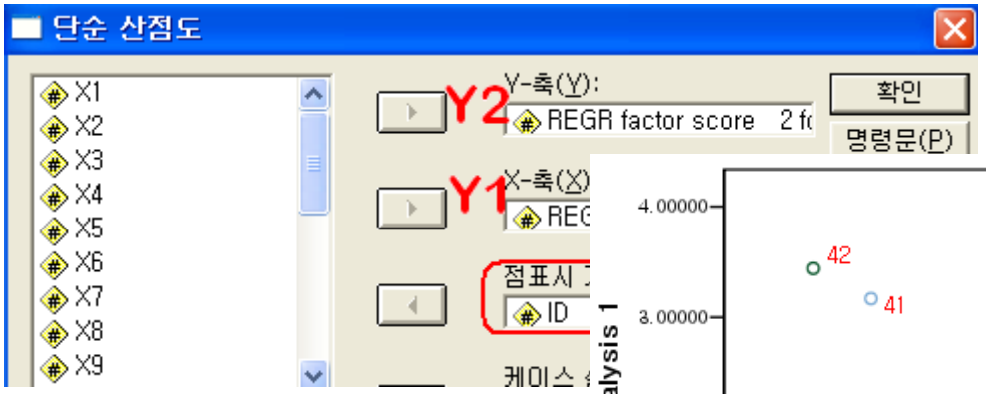
X10	X11	X12	X13	X14	X15	FAC1_1	FAC2_1	FAC3_1
8	9	7	5	7	10	.52765	-.08958	-.54000
9	9	8	8	8	10	1.24331	.09801	.02686

▪ 주성분 점수에 의한 개체 순위

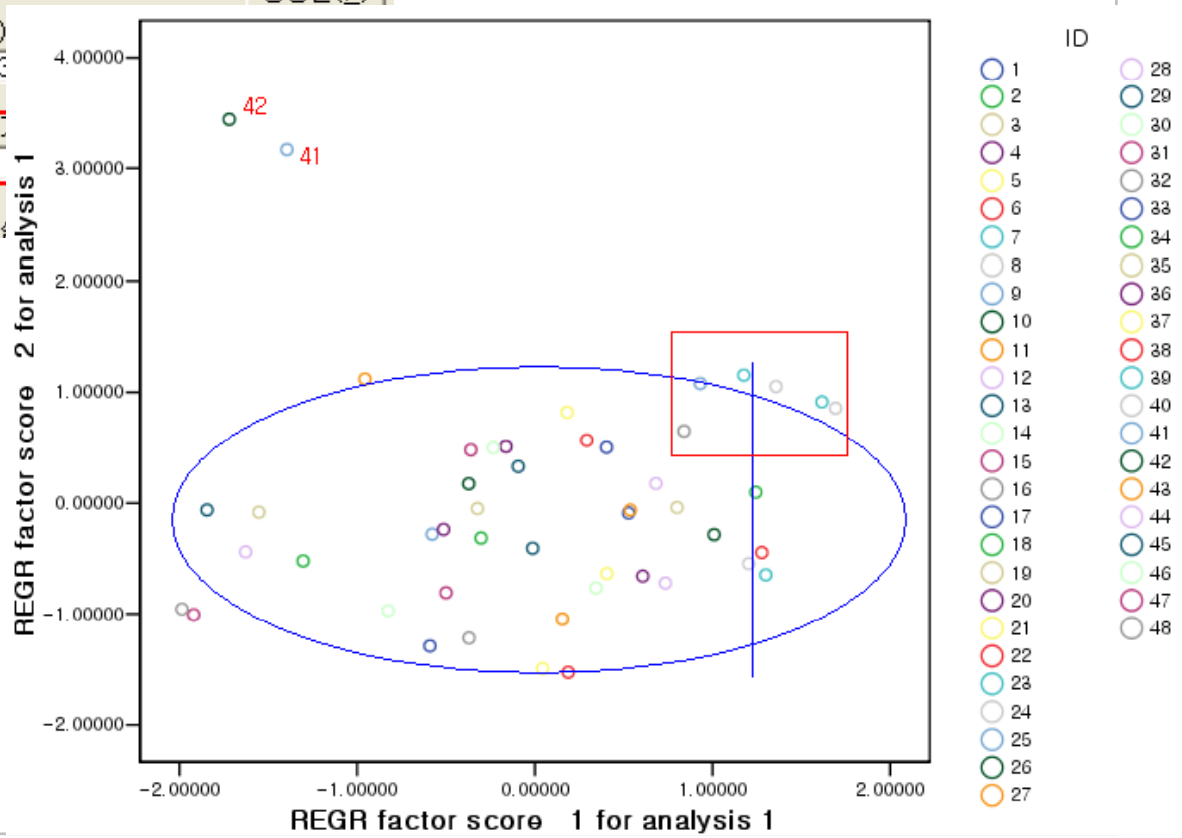


	ID	X1	X2	X3	X4	X5	X15	FAC1_1	FAC2_1	FAC3_1
1	40	10	6	9	10	9	10	1.69234	.85609	.57360
2	39	10	6	9	10	9	10	1.61581	.91339	.72387
3	8	9	9	9	8	9	10	1.35740	1.05203	.02561
4	23	7	10	7	9	9	8	1.30004	-.64651	.40179
5	22	9	8	7	8	9	8	1.27727	-.44525	.29810
6	2	9	10	5	8	10	10	1.24331	.09801	.02686
7	24	9	8	7	10	8	8	1.20384	-.54550	.86759
8	7	9	9	8	8	8	10	1.17611	1.15442	.09432

▪ 이상치 발견



산점도의 점들을 보면 제일 주성분의 설명력이 제이 주성분에 비해 크므로(고유치가 가장 크다) 타원의 현태가 나타난다. 제일 주성분(지적 능력)에 의해 지원자들 선발하면 수직선 오른편에 있는 6명을 선발하면 된다. ID 41, 42번 지원자는 제일 주성분은 낮고 제이 주성분은 매우 높아 이상치이다. 지적 능력은 낮으나 경험이 매우 많은 지원자이다.

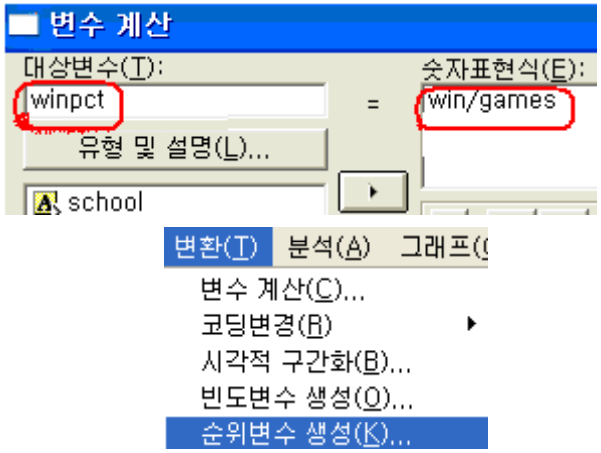


■ [http://lib.stat.cmu.edu/DASL]] . BIG8.SAV

▪ 학교명, 경기 수, 이긴 회수, Rushing공격 야드(RO_YDS), Rushing수비 야드, Passing공격 야드, Passing수비 야드

school	games	win	ro_yds	rd_yds	po_yds	pd_yds
Colorado	11	10.0	291.5	114.2	203.8	125.2
Iowa st.	11	.0	178.0	272.8	137.1	137.1
Kansas	11	6.0	247.1	171.2	140.9	135.8
Kansas st	11	9.0	125.6	167.5	237.6	94.3
Missouri	12	3.5	107.9	235.3	202.5	138.5
Nebraska	12	12.0	340.0	79.3	137.8	96.7
Oklahoma	11	6.0	182.2	148.5	173.9	107.5
Oklahoma	11	3.5	204.6	192.5	133.5	133.5

▪ 승률 계산 및 순위 부여



school	games	win	ro_yds	rd_yds	po_yds	pd_yds	winpct	Rwinpct
Colorado	11	10.0	292	114	204	125	.91	2.0
Iowa st.	11	.0	178	273	137	137	.00	8.0
Kansas	11	6.0	247	171	141	136	.55	4.5
Kansas st.	11	9.0	126	168	238	94.3	.82	3.0
Missouri	12	3.5	108	235	203	139	.29	7.0
Nebraska	12	12.0	340	79.3	138	96.7	1.00	1.0
Oklahoma	11	6.0	182	149	174	108	.55	4.5
Oklahoma st	11	3.5	205	193	134	134	.32	6.0

- 설명 변동 비율 및 부하

설명된 총분산

성분	초기 고유값			추출 제곱합 적재값		
	전체	% 분산	% 누적	전체	% 분산	% 누적
1	2.096	52.398	52.398	2.096	52.398	52.398
2	1.487	37.171	89.569	1.487	37.171	89.569
3	.362	9.062	98.631			
4	.055	1.369	100.000			

추출 방법: 주성분 분석.

성분점수 계수행렬

	성분	
	1	2
ro_yds	-.382	.367
rd_yds	.463	.058
po_yds	.024	-.627
pd_yds	.341	.376

- 제일 주성분은 수비 능력으로 볼 수 있다. 제이 주성분은 패싱 공격 능력이라 할 수 있다. 허용한 야드가 적을수록 수비 능력이 높은 것이므로 제일 주성분이 작은 값일수록 순위 높은 팀이다. 제이 주성분에서 패싱 야드 계수가 음이므로 이것 역시 낮은 학교가 순위가 높을 것이다.

- 주성분 순위 부여

school	gam	win	rod_	rdd_	pod_	pdi_	Rwinpct	FAC1_1	FAC2_1	RFAC1_1	RFAC2_1
Colorado	*	*	*	*	*	*	1	2.0	-.733	-.113	2.0	4.0
Iowa st.	*	0	*	*	*	*	0	8.0	1.169	.803	7.0	6.0
Kansas	*	6	*	*	*	*	1	4.5	.060	.940	5.0	8.0
Kansas st.	*	9	*	*	*	*	1	3.0	-.084	-1.983	4.0	1.0
Missouri	*	4	*	*	*	*	0	7.0	1.292	-.559	8.0	2.0
Nebraska	*	*	*	*	*	*	1	1.0	-1.786	.546	1.0	5.0
Oklahoma	*	6	*	*	*	*	1	4.5	-.294	-.469	3.0	3.0
Oklahoma st	*	4	*	*	*	*	0	6.0	.376	.835	6.0	7.0

- 이상치 발견하자.

개념

- 개체의 특성을 측정한 변수들 간에는 구조적 연관 관계(상관 관계)가 존재한다. ▷ 공분산 구조
- 구조적 연관 관계 크기(상관계수)에 따라 변수들을 그룹화
  - 요인 분석은 변수들의 내재된 상관 관계를 이용하여 요인을 구하고 이를 이용하여
  - (1)변수들을 분류하고 (변수 그룹에는 원 변수 일부만 포함되어 있다)
  - (2)그룹에 적절한 의미를 부여하는(그룹 이름 부여) 분석 방법이다.

역사

- 요인 분석(FA: Factor Analysis 혹은 인자 분석이라고도 함)은 사람의 지적 능력을 측정하고 이에 연관된 변수들을 이해하려는 노력의 일환으로 Galton(1888)에 의해 제안되었다.
- 수학적인 모형은 Spearman(1904 상관 계수 제안자)에 의해 발전
  - 상관계수만으로 과목의 구조적 특성 파악 난관, 과목의 구조를 설명하는 새로운 개념 factor 유도

	Classic	French	English	Math	Discover	Music
Classic	1	.83	.78	.7	.66	.63
French		1	.67	.67	.65	.57
English			1	.64	.54	.51
Math				1	.45	.51
Discover					1	.4
Music						1

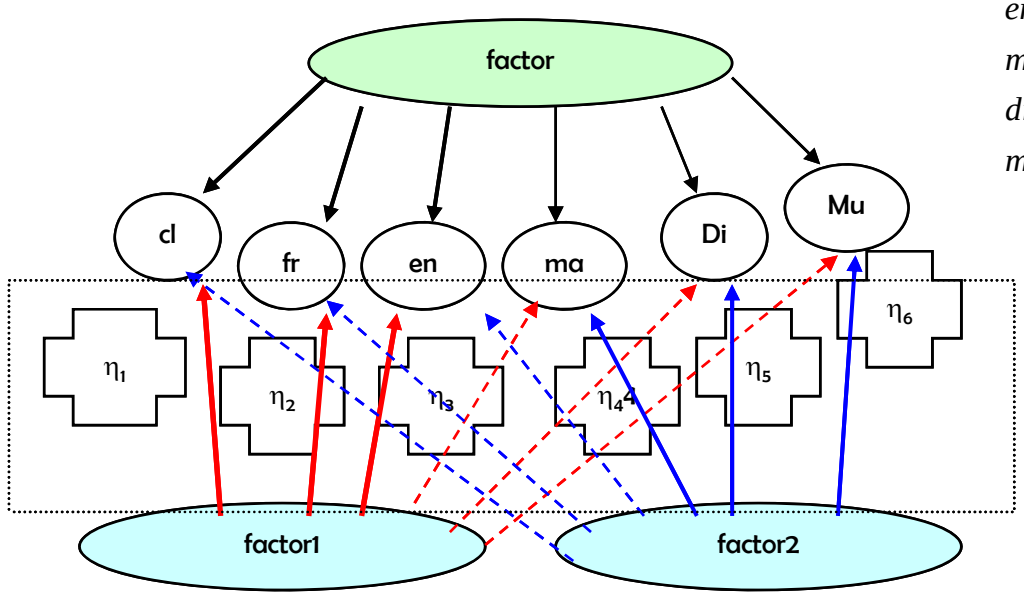
다변량 데이터

- 개체들의 특성을 측정한 (유사)변수들이 서로 상관 관계를 맺고 있어
- 직접적인 해석이 어렵고 여러 변수들간의 구조적 연관 관계를 파악하기 힘든 경우
- 요인(인자, factor)을 이용하여
  - 원변수의 내재된 구조적 관계를 이해한다.
  - 요인의 개수를 결정한다. 그룹의 크기
  - 변수 그룹의 이름 부여, 부하 행렬 이용

요인(factor)이란?

- Latent 변인, 공통 변인

$$classic = \lambda_1 f + \eta_1$$
$$french = \lambda_2 f + \eta_2$$
$$english = \lambda_3 f + \eta_3$$
$$math = \lambda_4 f + \eta_4$$
$$dis\,cover = \lambda_5 f + \eta_5$$
$$music = \lambda_6 f + \eta_6$$



요인분석과 상관분석(산점도 행렬)

- Football.xls
  - 변수: RO_YDS, RD_YDS, PO_YDS, PD_YDS, TO_YDS, TD_YDS
- 변수의 구조적 관계 상관관계에 따라 요인의 loading 값 크기가 결정된다.
- 변수의 그룹화는 상관관계 정도에 의존

상관: RO_YDS, RD_YDS, PO_YDS, PD_YDS,

	RO_YDS	RD_YDS	PO_YDS	PD_YDS	TO_YDS
RD_YDS	-0.738 0.037				
PO_YDS	-0.469 0.241	-0.078 0.854			
PD_YDS	-0.194 0.645	0.668 0.070	-0.324 0.433		
TO_YDS	0.840 0.009	-0.876 0.004	0.073 0.864	-0.444 0.271	
TD_YDS	-0.426 0.292	0.884 0.004	-0.233 0.579	0.907 0.002	-0.644 0.085

인자 분석: RO_YDS, RD_YI

상관 행렬에 대한 주성분 인자 분

비회전 인자 적재 및 공통성

변수	요인1	요인2	요인3
RO_YDS	-0.727	-0.674	-0.104
RD_YDS	0.978	0.046	0.051
PO_YDS	-0.096	0.874	-0.475
PD_RAT	0.760	-0.492	-0.380
TO_YDS	-0.890	-0.219	-0.380
TD_YDS	0.912	-0.300	-0.242
분산	3.6961	1.6013	0.5860
% Var	0.616	0.267	0.098

■유사점

- 변수들의 구조를 파악하기 위하여 공분산행렬(상관계수 행렬)를 이용한다.
- 새로운 변수가 만들어진다. 주성분점수, 요인점수

■차이점

- PCA는 원변수 X 차수 축약, FA는 원변수 X를 그룹화

주성분 분석	요인 분석
주성분은 원 변수의 직교 선형 결합으로 표현 Y=LX, L은 선형계수 행렬	인자들의 직교 선형 결합으로 원 변수들을 표현 X=LF, L은 부하행렬, F는 관측불가
주성분은 변수들의 변동을 설명한다.	요인은 변수들의 분산-공분산 구조 설명한다.
요인분석이나 주성분 분석의 L을 구하는 방법 유사하다. 공분산 행렬, 상관 행렬로부터 고유치 그에 대응하는 고유 벡터를 이용	
행렬 L은 변수의 개수 축약하는데 사용되며 는 주성분의 이름을 붙이는데 사용	행렬 L은 변수에 내재된 관계를 알아보는데 사용되며 는 변수들을 그룹화 하는데 사용한다.
적절한 주성분의 수를 구하고 주성분의 이름을 부여하고 주성분들간 산점도로 이상치 발견하거나 각 주성분 점수에 의해 개체 순위	적절한 인자의 수를 구하고 이를 이용하여 변수들을 그룹화 하고 그룹을 이용하여 변수에 내재된 관계를 알아본다.

Applicant.XLS

- 입사 지원자들에 대한 15개 항목 평가 주성분 분석 / 요인분석 결과
  - 주성분: 15개 항목 축약, 요인분석: 15개 항목 그룹화

주성분 분석: 이력서, 외모, 대

공분산 행렬에 대한 고유 분석

고유값	66.536	18.181	10.591	6.7
비율	0.543	0.148	0.086	0.0
누적	0.543	0.691	0.778	0.8

고유값	1.136	0.873	0.707	0.509
비율	0.009	0.007	0.006	0.004
누적	0.980	0.988	0.993	0.998

l_{ij}

변수	PC1	PC2	PC3
이력서	-0.149	0.371	0.200
외모	-0.132	-0.029	0.042
대학_학점	-0.030	0.102	-0.131
친밀감	-0.203	-0.093	0.620
자신감	-0.231	-0.236	-0.189
명석성	-0.337	-0.196	-0.125
친실성	-0.120	-0.301	0.447
겸손함	-0.379	-0.090	-0.282
추진력	-0.164	0.636	0.025
비전	-0.316	0.012	-0.113
장래성	-0.312	-0.122	-0.245
화합성	-0.339	-0.074	-0.050
업무적합성	-0.357	-0.025	0.041
마케팅	-0.226	-0.045	0.385
업무_파악	-0.274	0.471	0.017

주성분분석
공분산 행렬 사용

고유값 동일
요인/주성분 원변수 설명 정도 동일

요인 loading : 원변수 그룹에 사용
X=L(loading)F
주성분 계수: 주성분 이름 부여에 사용
Y=L(linear coefficient)X

$$f_{ij} = \sqrt{\lambda_i} l_{ij}$$

요인분석
요인추출: 주성분 이용
공분산 행렬 사용

인자 분석: 이력서, 외모, 대학_

공분산 행렬에 대한 주성분 인자 분석

비회전 인자 적재 및 공통성 f_{ij}

변수	요인1	요인2	요인3
이력서	-1.216	1.584	0.652
외모	-1.079	-0.125	0.136
대학_학점	-0.242	0.434	-0.426
친밀감	-1.657	-0.397	2.017
자신감	-1.888	-1.005	-0.616
명석성	-2.748	-0.836	-0.406
친실성	-0.981	-1.282	1.455
겸손함	-3.092	-0.384	-0.916
추진력	-1.338	2.713	0.081
비전	-2.578	0.053	-0.369
장래성	-2.546	-0.521	-0.796
화합성	-2.763	-0.317	-0.164
업무적합성	-2.913	-0.106	0.134
마케팅	-1.844	-0.191	1.254
업무_파악	-2.239	2.008	0.055
분산	66.536	18.181	10.591
% Var	0.543	0.148	0.086

■모형

$$\begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_p \end{pmatrix} = \begin{pmatrix} l_{11}\, l_{12} \cdots l_{1m} \\ l_{21}\, l_{22} \cdots l_{2m} \\ \cdots \\ l_{p1}\, l_{p2} \cdots l_{pm} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \cdots \\ f_m \end{pmatrix} + \begin{pmatrix} \eta_1 \\ \eta_2 \\ \cdots \\ \eta_p \end{pmatrix} \Leftrightarrow \underline{x} = L\,\underline{f}$$

- x_i는 i번째 원변수 ▷ x: 원변수 벡터
- f_i는 i번째 공통요인 (common factor) ▷ f: 요인 벡터
- l_{ij}는 i변수에 대한 j요인부하 (factor loading) ▷ L: 부하 행렬 (factor loading matrix)
- η_i는 i번째 원변수에 대한 오차항(공통인자가 설명하지 못하는 원인), 특정요인(specific factor)라 함

■가정

- f_i는 서로 독립이고 평균 0, 분산 1인 동일분포를 따른다. $\underline{f} \sim iid(\underline{0}, I)$
- η_i는 서로 독립이고 평균 0, 분산 w_i인 동일분포를 따른다. $\underline{\eta} \sim (\underline{0}, \psi = diag(\psi_1, \psi_2, .., \psi_p))$
- f_i와 η_i는 서로 독립이다. $Cov(\underline{\eta}, \underline{f}) = 0$

모형 해석

- 공통요인은 원변수 변동을 100% 설명한다.

$$\Sigma = Cov(\underline{x}) = Cov(L\underline{f} + \underline{\eta}) = LCov(\underline{f})L' + \psi = LL' + \psi \qquad \underline{\eta} \sim (\underline{0}, \psi = diag(\psi_1, \psi_2, .., \psi_p)$$

- i번째 원변수 변동=공통성(communality)+특정분산(specific variance), m<p

$$Var(x_i) = \sigma_{ii} = l_{i1}^2 + l_{i2}^2 + ... + l_{im}^2 + \Psi_i = \sum_{k=1}^{m} l_{ik}^2 + \psi_i$$

$$\begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_p \end{pmatrix} = \begin{pmatrix} l_{11}\, l_{12} \ldots l_{1m} \\ l_{21}\, l_{22} \ldots l_{2m} \\ \cdots \\ l_{p1}\, l_{p2} \ldots l_{pm} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \cdots \\ f_m \end{pmatrix} + \begin{pmatrix} \eta_1 \\ \eta_2 \\ \cdots \\ \eta_p \end{pmatrix}$$

  - 만약 m=p인 경우에는 특정분산=0이다.
  - 공통성은 요인들이 설명하는 원변수 변동

- 원변수 xi, xj의 공분산: $\mathrm{cov}(x_i, x_j) = \sum_{k=1}^{m} l_{ik} l_{jk}$
- 원변수 xi, 요인변수 fj의 공분산 $\mathrm{cov}(x_i, f_k) = l_{ik}$

요인 개수 m<p

- 차원 축소
  - 변수 그룹화
  - 일부 요인점수 사용 (고유치 1 이상, 80% 규칙)

요인방정식 해 구하기 (X=LF)

- principal factoring w/ or w/o iteration , Rao's canonical factoring, alpha factoring, image factoring, maximum likelihood, un-weighted least square factor analysis, Harris factoring
- Minitab: 주성분 방법(*), 최대우도 방법 제공

주성분 방법

- 원변수 공분산행렬(상관계수 행렬)로부터 고유치 고유벡터를 구한다. (주성분과 동일)

$$\Sigma = [\sqrt{\lambda_1}e_1 \mid \sqrt{\lambda_2}e_2 \mid \cdots \mid \sqrt{\lambda_p}e_p] \begin{bmatrix} \sqrt{\lambda_1}e_1' \\ \sqrt{\lambda_2}e_2' \\ \cdots \\ \sqrt{\lambda_p}e_p' \end{bmatrix} = LL' \Rightarrow \quad \Sigma = [\sqrt{\lambda_1}e_1 \mid \sqrt{\lambda_2}e_2 \mid \cdots \mid \sqrt{\lambda_m}e_m] \begin{bmatrix} \sqrt{\lambda_1}e_1' \\ \sqrt{\lambda_2}e_2' \\ \cdots \\ \sqrt{\lambda_m}e_m' \end{bmatrix} + \begin{bmatrix} \Psi_1 & 0 & \cdots & 0 \\ 0 & \Psi_1 & \cdots & 0 \\ --- & & & \\ 0 & \cdots & 0 & \Psi_p \end{bmatrix} = LL' + \Psi \qquad \Psi_i = s_{ii} - \sum_{i=1}^{m} l_{ij}^2$$

  - i번째 공통요인에 의해 원변수의 설명 변동 크기는 λ_i이다.
  - 요인과 주성분의 관계 $f_i = y_i\sqrt{\lambda_i}$

최대우도 추정방법

- 원변수가 다변량 정규분포 가정(강한 가정: 비율, 발생 회수, 소득, 자동차 가격 등 예외) 하에서 $\max_{L,\Psi} L(\underline{\mu}, \Sigma \mid \underline{x}) = L(\underline{\mu}, LL' + \Psi \mid \underline{x})$
  - L의 초기치로 다중 상관계수 제곱을 취하고 큰 공통성을 가진 변수에는 큰 가중치를 주게 되므로 공통성의 추정치가 1 이상이 되는 Heywood가 발생한다. 이 상황에서는 Ψ의 추정치가 음이 된다. (사용불가)

부하(l_{ij}, loading) 의미

- 각 요인(원변수 내재된 관계의 공통 부분)이 원변수에 미치는 영향력 크기
- 각 요인에서 부하 절대값이 큰 것들만 선택하여 변수들을 그룹화 한다.
  - 그룹화: 그룹 내 변수들의 평균 값 계산
  - 부하 값이 음인 경우는 반대 개념: 반대 부호는 (-)를 붙여준 후 평균 계산
- 요인의 개수는 변수 그룹 개수를 의미한다.
- 공통성 (communality)
  - 요인들이 설명하는 원변수 변동 부분 $Var(x_i) = \sigma_{ii} = l_{i1}^2 + l_{i2}^2 + \ldots + l_{im}^2 + \Psi_i = \sum_{k=1}^{m} l_{ik}^2 + \psi_i$
  - 원변수 분산=공통성+특정분산

$$\begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_p \end{pmatrix} = \begin{pmatrix} l_{11}\, l_{12} \ldots l_{1m} \\ l_{21}\, l_{22} \ldots l_{2m} \\ \cdots \\ l_{p1}\, l_{p2} \ldots l_{pm} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \cdots \\ f_m \end{pmatrix} + \begin{pmatrix} \eta_1 \\ \eta_2 \\ \cdots \\ \eta_p \end{pmatrix}$$

factor1	factor2	Error
Common factors		

요인 개수 구하기

- (1)trivial한 요인은 제외하자. 원 변수 1-2개에만 부하 값이 큰 요인은 제외하자. 이 요인에 의해 묶을 수 있는 변수는 1-2개이므로 그룹의 의미가 없기 때문이다.
- (2)Kaiser 판단(가장 많이 이용): 변수들의 상관 관계가 0이면 (즉, 상관계수 행렬 R은 항등 행렬 I) 원 변수의 개수와 주성분의 개수가 같아지고 주성분의 분산은 모두 1이고 각 주성분이 가지는 분산 평균도 1이다. 이를 이용하여 상관계수 행렬로부터 구한 고유치가 평균인 1이상인 되어야 한다는 판단 하에 고유치가 1 이상인 것만으로 요인의 개수 결정
- Scree Plot 사용: 주성분과 동일, 80% 규칙 사용 가능 (% Var: 분산 변동 설명비율)
- 요인 개수 강제 결정: 그룹의 개수를 미리 결정하거나, 원변수를 어느 한 그룹에 반드시 속하게 할 때 사용

개념

- 요인의 개수 m<p인 경우 부하 행렬의 해는 많은 해가 존재.
- 요인 부하 값에 의해 원 변수를 그룹화 한다, but
  - (1)요인의 복합성: 하나의 원 변수에 부하 값이 큰 요인이 2개 이상 존재하거나
  - (2)인자의 크기가 0을 중심으로 ±의 작은 값이 있는 경우 부하 값으로 변수를 그룹화 하는 것은 불가능하다.

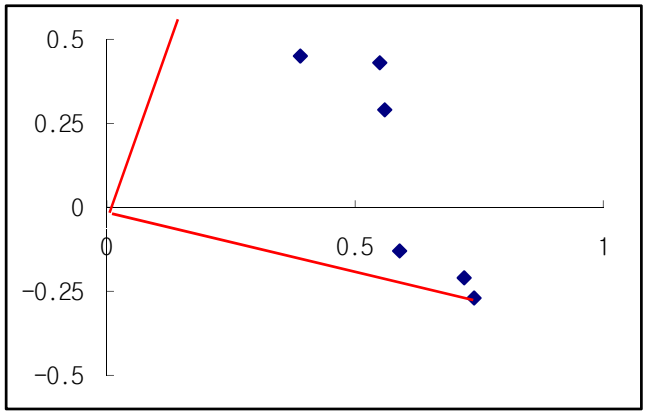
정의

- 각 요인이 상대적으로 큰 부하 값을 갖도록 요인을 회전(rotate)
- 방법: QUARTIMAX rotation, OBLIQUE rotation, PROMAX rotation, VARIMAX(*)
- VARIMAX 방법: (Kaiser 제안) 간단한 구조 측정치로 부하행렬의 분산을 최대화 하는 회전 방법

예제

변수	요인1	요인2
X1	0.55	0.43
X2	0.56	0.29
X3	0.39	0.45
X4	0.74	-0.27
X5	0.72	-0.21
X6	0.59	-0.13

$$\mathrm{cov}(x_i, x_j) = \sum_{k=1}^{m} l_{ik} l_{jk}$$



요인점수 F 의미

- 원변수에 내재된 공통성 측정
- 요인들은 서로 독립이다.
- 제일 요인이 원변수 구조를 가장 많이 설명, 제이 요인...
- 주성분 점수와 비교
  - 주성분은 원변수의 선형결합, 원변수에 의해 만들어짐
  - 요인점수는 원변수의 내재된 공통 개념

L의 의미

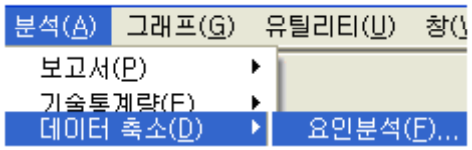
- 주성분 (Y=LX): 원변수가 주성분에 미치는 영향, 주성분 이름 부여
- 요인(X=LF): 요인이 원변수에 미치는 영향, 원변수 그룹화에 활용, 역으로 요인점수 이름 부여에 사용하기도 한다.

요인점수 구하기 ((X=LF+η)

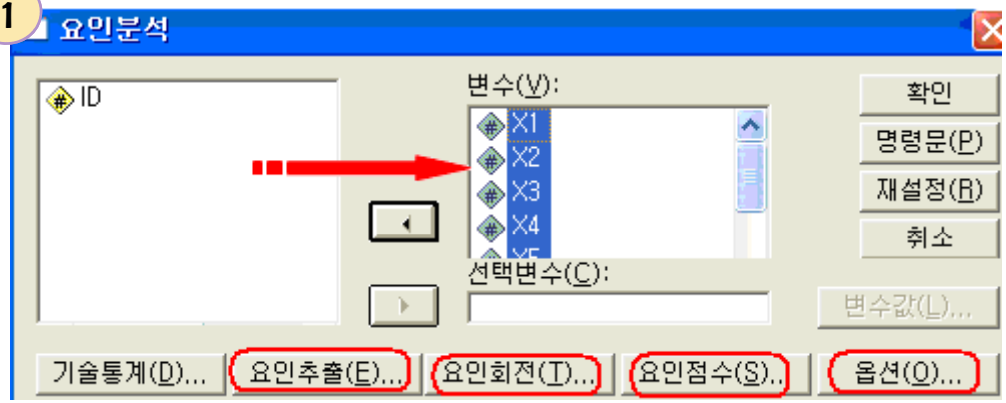
- 관측치 X=> L 계산 , 그러나 unknown η
  - Bartlett's Method (Weighted Least Square Method) $\underline{z}_r = (\underline{x}_r - \underline{\mu}) \Rightarrow \underline{f}_r = (\hat{L}\hat{\psi}^{-1}\hat{L})^{-1}\hat{L}\hat{\psi}^{-1}\underline{z}_r$
  - Thompson's Method (Regression Method)

$$\begin{bmatrix} \underline{z} \\ \underline{f} \end{bmatrix} \sim N\left(\begin{bmatrix} \underline{0} \\ \underline{0} \end{bmatrix}, \begin{bmatrix} P & L \\ L' & I \end{bmatrix} \right) \Rightarrow E(\underline{f} \mid \underline{z}) = L'P^{-1}\underline{z} \Rightarrow \underline{f}_r = L'R^{-1}\underline{z}_r$$

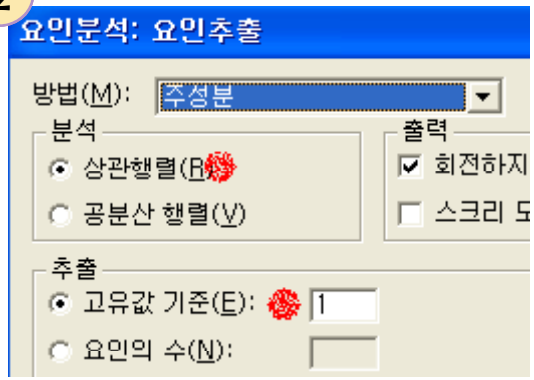
Applicant.txt 예제 계속



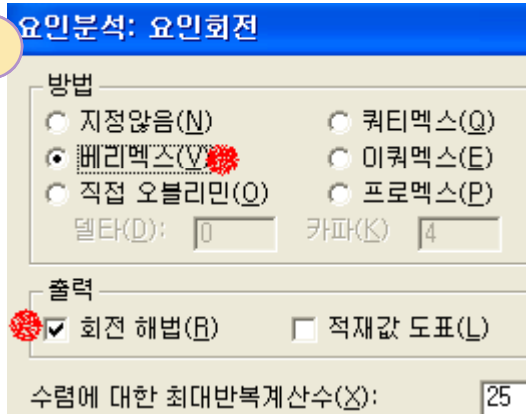
1



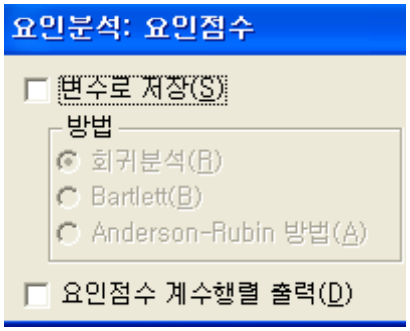
2

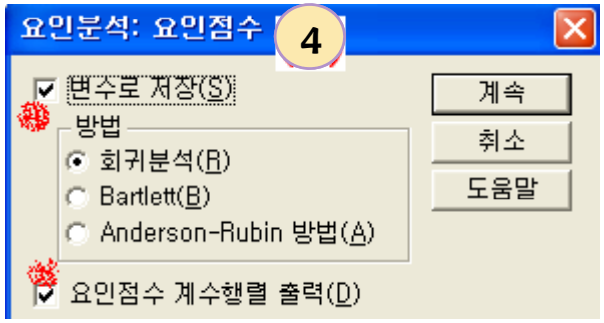


3



요인 점수 저장?





- 변수의 단위와 속성이 유사하면 변수 그룹 내의 변수 평균으로 요인점수 사용
- 변수 단위나 속성이 다르면 요인점수(이것이 주성분 점수 개념) 구하기에 의해 계산된 값 사용

변동 설명 비율

- 상관 계수 행렬이 사용되었으므로 고유치의 합은 원 변수 개수 15이다.
- 고유치가 1이상 공통 요인의 원 변수 변동의 설명력은 80% 이상이다.

설명된 총분산

성분	초기 고유값		
	전체	% 분산	% 누적
1	7.514	50.092	50.092
2	2.056	13.709	63.801
3	1.456	9.705	73.506
4	1.198	7.986	81.492
5	.739	4.928	86.420
6	.495	3.297	89.717
7	.351	2.342	92.059
8	.310	2.066	94.125
9	.257	1.713	95.838
10	.185	1.233	97.071
11	.153	1.018	98.088
12	.098	.650	98.739
13	.089	.592	99.331
14	.065	.431	99.762
15	.036	.238	100.000

추출 방법: 주성분 분석.

부하 활용

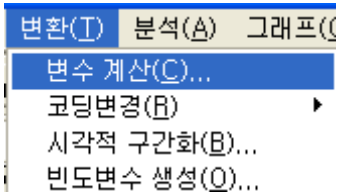
회전된 성분행렬[a]

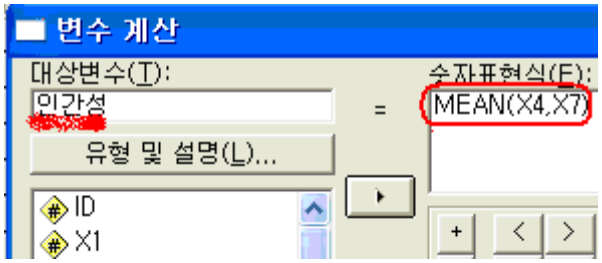
	성분			
	1	2	3	4
X11	.918	.159	.100	-.041
X5	.916	-.107	.163	-.065
X8	.910	.223	.103	-.041
X6	.863	.097	.255	.002
X12	.811	.255	.331	.143
X10	.800	.349	.156	-.052
X13	.747	.326	.413	.224
X2	.440	.151	.399	.227
X9	.087	.851	-.055	.211
X1	.116	.830	.109	-.136
X15	.383	.797	.076	.084
X4	.220	.245	.871	-.081
X7	.219	-.242	.863	.001
X14	.440	.363	.534	-.524
X3	.064	.128	.007	.928

요인추출 방법: 주성분 분석.
회전 방법: Kaiser 정규화가 있는 베리멕스.

그룹1: (자신감 x5), (명석 x6), (마케팅 능력 x8), (추진력 x10), (야망 x11), (개념 파악 능력 x12) , (장래성 x13)
그룹 2: (이력서 x1), (경험 x9), (업무 적합성 x15)
그룹 3: (친밀감 x4), (진실 x7)

결과 활용





ID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	인간성
1	6	7	2	5	8	7	8	8	3	8	9	7	5	7	10	6.50
2	9	10	5	8	10	9	9	10	5	9	9	8	8	8	10	8.50
3	7	8	3	6	9	8	9	7	4	9	9	8	6	8	10	7.50
4	5	6	8	5	6	5	9	2	8	4	5	8	7	6	5	7.00
5	6	8	8	8	4	4	9	5	8	5	5	8	8	7	7	8.50

- 측정 변수 단위와 속성이 동일하면 변수의 평균 사용
- 설문지 리커트 척도 하위문항 구성에 사용
  - exploratory factor analysis
  - confirmatory factor analysis (LISREL)

POLICE.SAV

- 경찰에 지원한 50명의 신체적 특성 15개를 측정하였다. [Applied Multivariate Methods for Data Analysts, Dallas E. Johnson, 1998, p160]
  - ID: 지원자 번호/REACT: 시각적 자극에 대한 반응 시간/HEIGHT (cm) / WEIGHT (kg)
  - SHLDR: 어깨 넓이(cm) / PELVIC: 골반 넓이(cm) / CHEST: 가슴 넓이(cm)
  - THIGH: 허벅지 피부 두께 (mm) / PULSE: 맥박 / DIAST: 심장 혈압 / CHNUP: 턱걸이 회수
  - BREATH: 폐활량 (liter) / RECVR: 런닝머신에서 제자리 달리고 5분 후 맥박
  - SPEED: 런닝머신에서 제자리 달리기 최대 속력
  - ENDUR: 런닝머신에서 달릴 수 있는 최대 시간(분) / FAT: 비만도

ID	react	height	weight	shldr	pelviv	chest	thigh	pulse	diast	chnup	breath	recvr	speed	endur	fat
1.0	.3	180	74.2	41.7	27.3	82.4	19.0	64	64	2	158	108	5.5	4.0	12
2.0	.3	176	62.0	37.5	29.1	84.1	5.5	88	78	20	166	108	5.5	4.0	3.1
3.0	.3	166	73.0	39.4	26.8	88.1	22.0	100	88	7	167	116	5.5	4.0	17
4.0	.3	174	85.9	41.2	27.6	97.6	19.5	64	62	4	220	120	5.5	4.0	20

brief solution

설명된 총분산

성분	초기 고유값		
	전체	% 분산	% 누적
1	5.219	34.790	34.790
2	2.407	16.045	50.835
3	1.313	8.751	59.587
4	1.231	8.207	67.794
5	1.204	8.026	75.819
6	.848	5.653	81.472
7	.705	4.698	86.170
8	.578	3.856	90.027

회전된 성분행렬[a]

	성분				
	1	2	3	4	5
fat	.898	.304	-.017	.056	.046
thigh	.865	.074	.111	-.028	.057
chnup	-.830	-.106	.004	.074	-.146
chest	.607	.572	-.143	.118	-.178
endur	-.390	-.265	-.167	.369	.016
height	.115	.824	-.099	-.207	.295
shldr	.146	.821	-.043	-.132	-.170
pelviv	.161	.795	-.239	.268	-.105
weight	.653	.685	-.173	.025	-.041
breath	.191	.607	.205	-.330	.307
recvr	.107	-.045	.884	.012	-.197
pulse	-.144	-.130	.785	.117	.196
speed	-.383	.169	-.493	-.463	-.067
diast	-.016	.010	.185	.868	.093
react	.119	-.004	-.007	.113	.935

요인추출 방법: 주성분 분석.
회전 방법: Kaiser 정규화가 있는 베리멕스.

- 변수의 측정 단위와 속성이 다르므로 요인점수 활용하자.

Individual Directed Technique

- 측정 변수(항목)에 의한 개체 분류
  - 분류되어 있는 집단간의 차이를 의미 있게 설명해 줄 수 있는 독립변수들을 찾아내어
  - 변수의 선형결합으로 판별식(Discriminant function) 을 만들어 낸다.
  - 이 판별식을 이용하여 분류하고자 하는 개체의 집단을 판별

데이터 유형

- 집단변수: 범주형 혹은 이진형
- 판별 변수: 측정형(등간 척도 포함)

사례

- 6개월 내 TU 해지 고객 판별 변수 및 판별함수 유도
- SKT/KT/LGT 가입 고객 판별 변수 및 판별함수 유도
- 서비스 이용 불만고객 성향 분석

Variable Directed Techniques
- 변수 축약: 주성분 점수
- 유사 변수그룹: 요인분석, 요인점수

- 개체분류
  - 군집분석
  - 판별분석

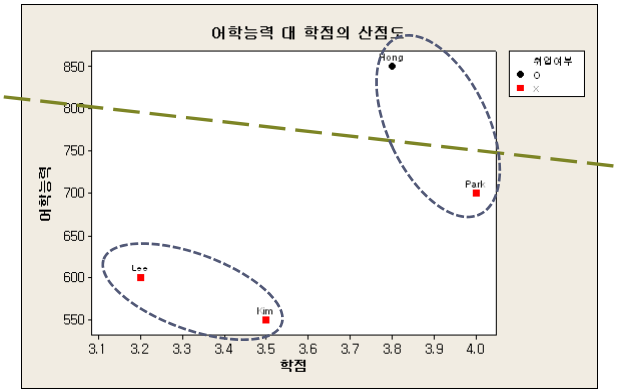
이름	취업 여부	어학 능력	학점	봉사활동
Kim	X	550	3.5	12 months
Lee	X	600	3.2	6 m
Park	X	700	4.0	0 m
Hong	O	850	3.8	24 m

주성분 점수나 요인점수 이용 개체 판별?

- 집단에 따른 주성분점수 차이 분석 ▷집단 성향 T-검정(이진형 집단), 분산분석(3 집단 이상)

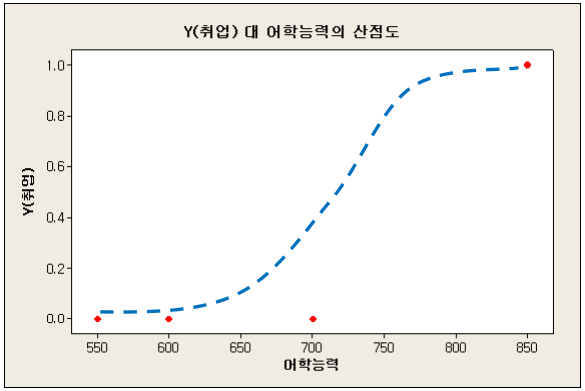
Clustering analysis

- (유사) 개체를 분류 (grouping)
- (상이)데이터에는 집단을 구별하는 변수 없음 ▷ 개체의 유사성(similarity)에 의해 개체 분류



Regression analysis

- (유사)
  - 집단 변수가 이진형 혹은 순서형 ▷ 종속변수, Logistic Regression
  - 판별 변수와 회귀분석 독립변수 집단 차이 설명
- (상이)
  - 판별분석은 집단이 범주형인 경우에도 가능
  - 집단을 구별하는 판별식 유도(집단 분류), 회귀분석은 연결함수 이용 선형모형화 (집단 소속 예측 확률))

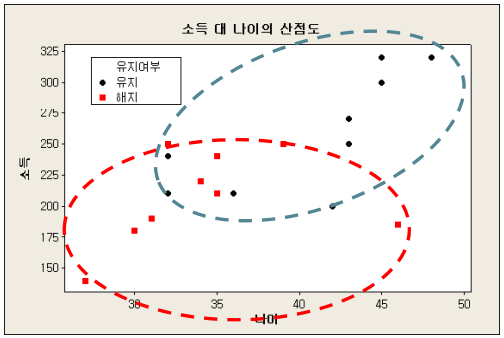


판별함수 (discriminant function)

- R=f(X1, X2, ..., Xp): 개체의 집단을 판별하는데 사용되는 판별변수의 함수

판별함수 찾기

- 집단 내 분산에 비해 집단간 분산의 차이를 최대화하는 독립변수의 함수를 찾는다.



판별함수 개수

- Min(집단 개수-1, 판별변수 개수)

데이터 크기

- 관측치(개체)의 개수(데이터의 크기, 표본 크기)가 판별변수 개수의 20배 이상
- 집단의 각 범주에 최소한 20개의 관측치가 요구
- 위의 조건을 충족시키지 못하면 분석결과는 불안정(판별식을 구성하는 각 독립변수와 전체 판별식의 설명력과 예측력을 신뢰할 수 없다는 의미)해 짐

판별함수 집단이 2개(k=1집단, 2집단) 인 경우, 판별변수 X1, X2, ..., Xp

Z: 판별점수, ai는 판별계수 $Z = a_1 X_1 + a_2 X_2 + ... + a_p X_p$

Fisher's Linear Discriminant Function (선형함수)

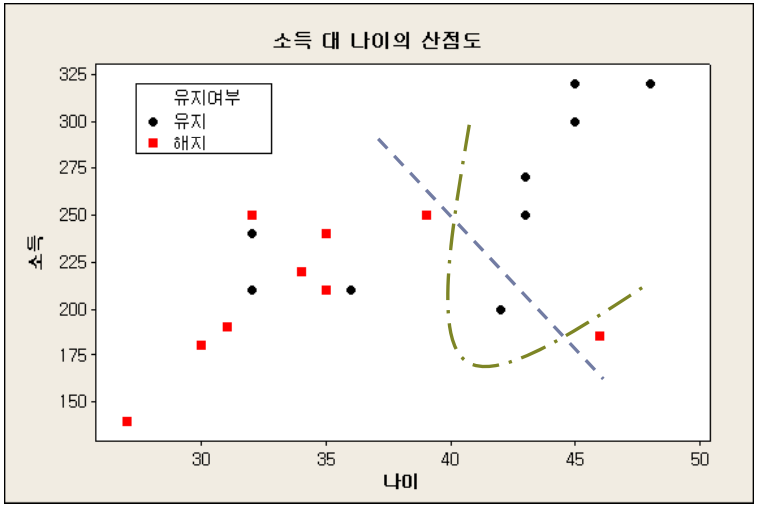
- j번째 개체: $(\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}\underline{x}_j - k > 0$ 집단 1에 분류, 그렇지 않으면 집단2에 분류 $k = (1/2)(\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$
- 두 집단의 분산이 같다는 가정, 판별식에 사용되는 개념은 Mahalanobis 거리

이차함수

- 두 집단의 분산이 같지 않을 때
- 거리의 제곱의 함수가 k와는 달리 이차 형식
- 선형함수에 비해 경계의 유연성은 있으나 항상 좋은 것은 아님

예제 데이터 서비스.MTW

- OO 서비스 해지여부를 판별하는 변수로 나이, 소득



오분류

- 판별함수 신뢰 정도 평가하는데 사용

분류집단 / 원 집단	집단1	집단2
집단1	정분류	오분류
집단2	오분류	정분류

오분류율 (misclassification ratio)

- (오분류 개체 수)/ (전체 개체 수) * 100
- 정분류율(=1-오분류율): 회귀분석의 결정계수 R^2 개념

오분류율 추정 방법

- Re-substitution 규칙
  - 모든 개체 사용하여 판별식을 구하고, 이를 이용하여 오분류 비율 계산
  - 간편하나 정분류율이 과대 추정 가능
- Cross-validation 방법
  - 개체 제외하고 판별식을 구하여 제외한 개체의 집단을 분류한다. 이 작업을 반복한다.
  - 가장 많이 사용
- 테스트 데이터 이용
  - 데이터를 이분하여, 한 데이터는 판별식(60~70%) 추정, 다른 데이터(40~30%)는 오분류율 계산에 사용
  - 가장 정확한 오분류 계산, 어느 정도 대용량 데이터 확보 필요 (data mining에서)

■비용함수

- 오분류에 의한 비용함수 고려하여 판별식 선택
- 비용함수 선택
  - Equal Cost function (균등비용함수)
  - Ratio cost function (비례비용함수)
- 비용함수 고려 모형 복잡하므로 ECF 사용하여 오분류 표를 얻은 후 비용을 사후적 고려하는 것이 편리
- Minitab은 비용함수 고려 옵션 없음

$$k_i^* = 1/2(\underline{x}_0 - \underline{\mu}_i)'\Sigma^{-1}(\underline{x}_0 - \underline{\mu}_i) - \ln(p_i^*)$$

$$p_1^* = \frac{p_1 C(2|1)}{p_1 C(2|1) + p_2 C(1|2)} \qquad p_2^* = \frac{p_2 C(1|2)}{p_1 C(2|1) + p_2 C(1|2)}$$

- 환자 마취 여부 판별
  - 판별식 1 사용이 적절

판별식1 ▷	마취 가능	마취 위험	판별식2 ▷	마취 가능	마취 위험
마취 가능	95	10	마취 가능	90	5
마취 위험	5	90	마취 위험	10	90

■사전 확률 개념

- 표본 데이터의 비율이 모비율 집단 구성 비율과 현저히 다를 때
- 집단 비율에 대한 사전 정보를 판별식에 이용

$$d_i^* = 1/2(\underline{x}_0 - \underline{\mu}_i)'\Sigma^{-1}(\underline{x}_0 - \underline{\mu}_i) - \ln(p_i^*)$$

$$p_1^* = \frac{p_1 C(2|1)}{p_1 C(2|1) + p_2 C(1|2)} \qquad p_2^* = \frac{p_2 C(1|2)}{p_1 C(2|1) + p_2 C(1|2)}$$

□SPSS는 사전확률을 설정할 수 있는 옵션은 없고 동일하게 설정하거나 표본의 크기 비율을 사용하는 옵션만 제공하고 있다. 비용함수에 대한 옵션은 없다.

판별변수 선택 개념

- 선택변수 모두 판별 능력이 있나?
- 어떤 변수의 판별 능력이 가장 큰가?
- (logic) 집단을 잘 분류한다? 집단 간 판별변수의 평균 차이 크다.
- (예제) (학점, 어학능력, 어학 연수기간)에 따른 취업집단 판별

선택 필요 이유 (parsimony rule)

- 측정 오류 발생 가능성이 적고
- 새로운 개체 판별을 위해 측정해야 하는 변수 수가 적어 효율적.

선택 방법

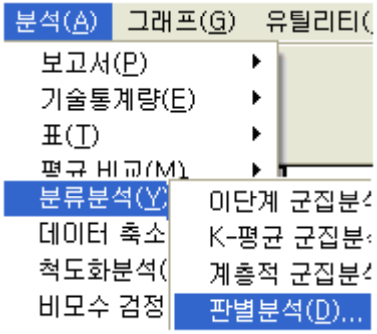
- 분산분석 및 공분산분석 개념
  - 분산분석에 의해 F값이 가장 큰 판별변수 선택 (예: 어학능력)
  - 선택된 판별변수(어학능력)를 공변량(covariate)으로 하여 공분산분석(ANOCOVA)으로 (학점, 연수기간)을 판별변수 선택
- Forward, Stepwise, Backward 방법

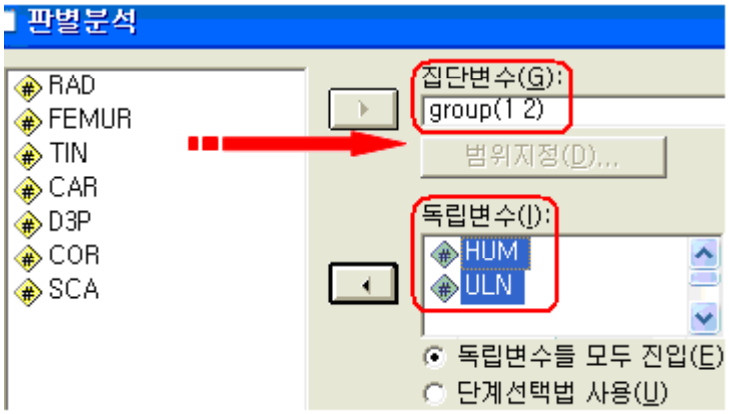
TURKEY.SAV

- 미국 Kansas 주립대학 Dr. Michael Finnegan 교수는 야생 칠면조와 사육 칠면조를 구별하기 위하여 수컷 칠면조 82마리에 대해 9개 항목을 조사하였다.
- [Applied Multivariate Methods for Data Analysts, Dallas E. Johnson, 1998]
  - ID: 칠면조 id HUM: 상완골 길이 RAD: 요골 길이
  - ULN: 척골) 길이 FEMUR: 대퇴골 길이 TIN: 경골 길이
  - CAR: carp metacarpus 길이 D3P: 지골까지 길이 COR: 오탁상 길이
  - SCA: 견갑골 길이 TYPE: 칠면조 종류 야생(WILD) 사육(DOMESTIC)

ID	HUM	RAD	ULN	FEMUR	TIN	CAR	D3P	COR	SCA	TYPE
B710	153	140	147	142	151	817	305	102	128	WILD
B790	156	137	151	146	155	814	305	111	137	WILD
B791	.	132	148	138	145	775	.	106	128	WILD
B795	151	134	151	144	.	789	292	116	126	WILD
B819	158	135	151	146	152	790	289	111	125	WILD

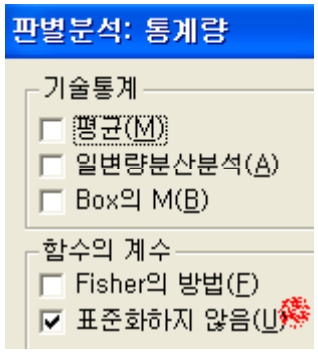
■메뉴





- 집단변수(그룹)가 숫자형으로 지정되어 있어야 한다.
- 변환(-T)➔코딩변경(R)을 통하여 WILD=1, DOMESTIC=2로 변환하였다
- 유의한 판별 변수 선택 방법으로 단계선택 방법 제공

▸다음 슬라이드 집단 분산 동일성 검정을 위한 옵션

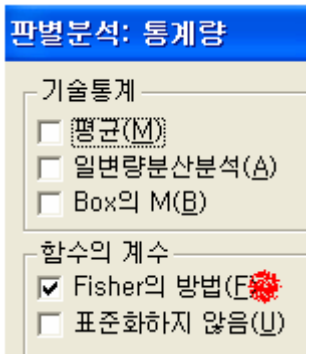


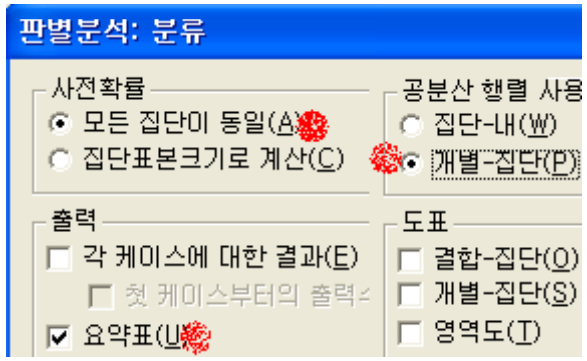
검정 결과

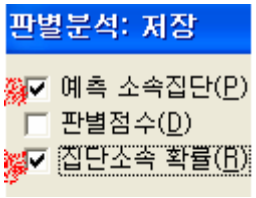
Box의 M		.137
F	근사법	.133
	자유도1	1
	자유도2	3603.527
	유의확률	.715

정준 판별 함수의 모집단 공분산행렬이 동일하다는 영가설을 검정합니다.

메뉴 cont.







- 함수의 계수는 판별에 사용되는 식을 얻기 위한 것이다. 표준화는 원 변수를 표준화 했을 때 식의 계수이다. 일반적으로 분석이나 해석에 사용하지 않으므로 선택할 필요는 없다.
- 사전확률에서 표본이 모집단의 비율에 맞추어 층화추출(stratified sampling) 되었다면 "집단 표본크기로 계산" 옵션을 선택하면 된다.
- 공분산 행렬은 사용된 분류 변수의 단위가 비슷하면 통합 분산(pooled)을 이용하자. 다르다면 집단 내(within) 분산 사용하면 되고... 등분간 검정 후에 결정하자.
- "요약표"는 오분류 표 출력하라는 옵션이다.
- "집단소속 확률"은 개체가 집단에 소속될 확률을 데이터에 저장하는 옵션이다. 개체는 소속 확률이 가장 큰 집단에 소속되는데 이에 대한 정보는 "예측 소속집단"을 선택하면 데이터에 변수로 저장된다.

■사전 확률 및 판별식 계수

집단에 대한 사전확률

group	사전확률	분석에 사용된 케이스	
		가중되지 않음	가중됨
1	.500	17	17.000
2	.500	20	20.000
합계	1.000	37	37.000

분류 함수 계수

	group	
	1	2
HUM	3.401	3.325
ULN	3.481	3.212
(상수)	-520.852	-470.190

Fisher의 선형 판별함수

$$Y_{야생} = -520.85 + 3.4 * HUM + 3.48 * ULN$$

$$Y_{사육} = -470.19 + 3.33 * HUM + 3.21 * ULN$$

■오분류 표

분류결과[a]

		group	예측 소속집단		전체
			1	2	
원래값	빈도	1	14	3	17
		2	5	15	20
	%	1	82.4	17.6	100.0
		2	25.0	75.0	100.0

a. 원래의 집단 케이스 중 78.4%이(가) 올바로 분류되었습니다.

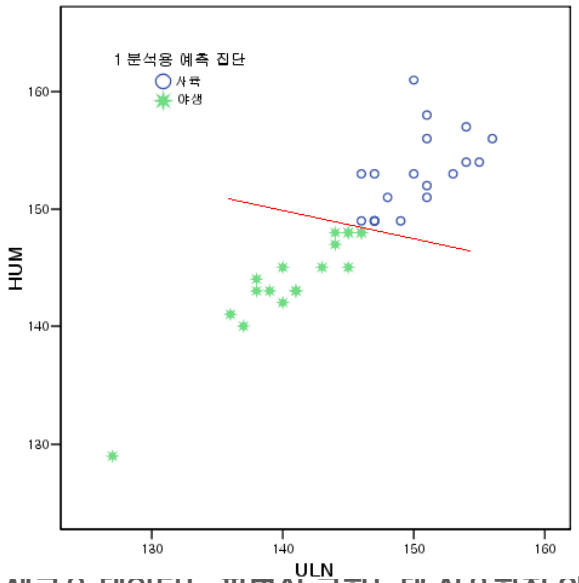
- 위의 판별식에 의해 GROUP=1(야생)이 사육으로 잘못 분류된 칠면조는 3마리(오분류 비율 17.6%), 사육인데 야생으로 분류한 칠면조는 5마리(오분류 비율 26.3%)이다. 전체 오분류 비율은 21.6%이다. 어느 칠면조가 오분류 되었는지는 데이터의 집단 분류 결과를 보면 된다.

판별 결과 보기

- 주성분 분석 이용하여 변수를 축약한 후 산점도
- 옆 그림은 판별변수 중 판별 능력이 가장 큰 두 변수만 사용

- 새로운 개체

ID	HUM	RAD	ULN	FEMUR	TIN	CAR	D3P	COR	SCA	TYPE	group
L770	144	129	138	130	134	790	300	97	118	DOM	2
L774	143	128	141	130	137	800	300	98	123	DOM	2
새1	145	.	150	.	.	.	.	.	.		.
새2	150	.	145	.	.	.	.	.	.		.



- 데이터 제일 아래 새로운 개체 2개를 다음과 같이 입력한 후 판별분석을 실시하면 새로운 데이터는 판별식 구하는데 사용하지 않고 새로운 데이터가 각 집단에 속할 확률과 어느 집단에 속하는지 출력된다.

분류결과[a]

		group	예측 소속집단 1	예측 소속집단 2	전체
원래값	빈도	1	14	3	17
		2	5	15	20
		집단화되지 않은 케이스	1	1	2
	%	1	82.4	17.6	100.0
		2	25.0	75.0	100.0
		집단화되지 않은 케이스	50.0	50.0	100.0

a. 원래의 집단 케이스 중 78.4%이(가) 올바로 분류되었습니다.

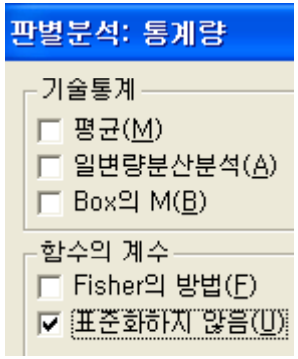
ID	HUM	RAD	ULN	FEMUR	TIN	CAR	D3P	COR	SCA	TYPE	group	Dis_1	Dis1_1	Dis2_1
L770	144	*	*	*	*	*	*	*	*	D	2	2	.06099	.93901
L774	143	*	*	*	*	*	*	*	*	D	2	2	.13208	.86792
새1	145	.	*	.	.	.	.	.	.		.	1	.69465	.30535
새2	150	.	*	.	.	.	.	.	.		.	2	.46490	.53510

WHEAT.SAV

- 밀(Wheat) 종류에는 Arthur종(soft한 밀)과 Arkan종(hard 밀)이 있고 Group 1, 2과 Group 3, 4는 서로 다른 지역이다. 그러므로 4개의 집단이 존재한다. 밀에 대해 다음 항목의 길이를 조사하였다. n=172
- 밀의 오른쪽(Right) 면에서 면적(Area R1), 원주(Perimeter R2), 길이(Length R3), 폭(breadth R4) 그리고 아래쪽(down)에서 면적(Area D1), 원주(Perimeter D2), 길이(Length D3), 폭(breadth D4)을 조사하였다.
- [Applied Multivariate Methods for Data Analysts, Dallas E. Johnson, 1998]

TYPE	R1	R2	R3	R4	D1	D2	D3	D4
1	54.45	219.0	89.00	43.0	56.60	226.0	89.00	47.0
1	55.15	221.0	91.00	46.0	56.26	224.0	91.00	46.0
1	53.92	223.0	90.00	44.0	55.09	223.0	91.00	44.0
1	52.23	212.0	87.00	41.0	53.54	215.0	88.00	44.0

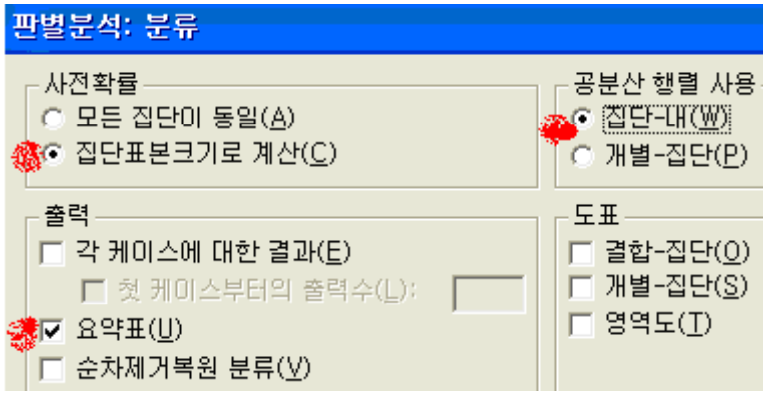
■집단 등분산 검정

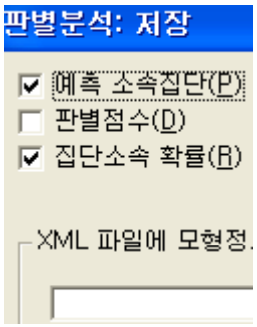


검정 결과

Box의 M		69.195
F	근사법	3.716
	자유도1	18
	자유도2	81222.713
	유의확률	.000

▪ 귀무가설이 기각되어 등분산 가정이 무너지므로 집단-내 공분산을 사용한다.





오분류 표

분류결과[a]

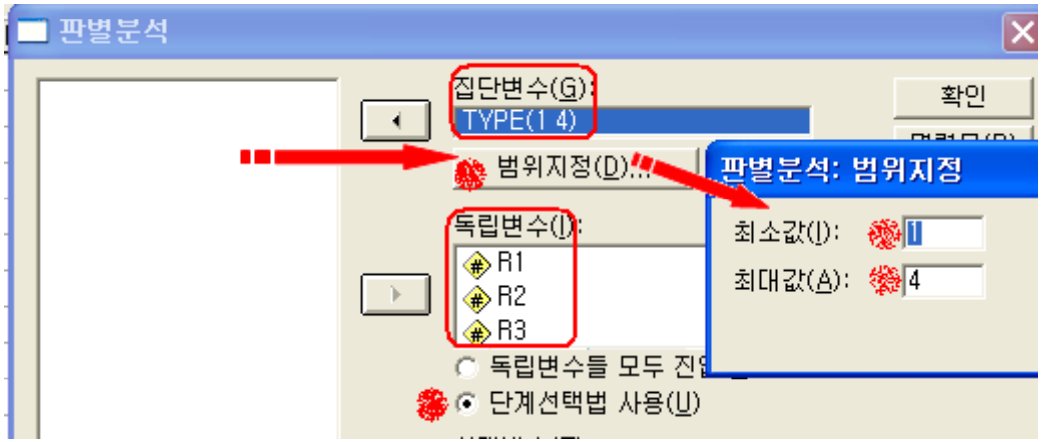
		TYPE	예측 소속집단 1	2	3	4	전체
원래값	빈도	1	23	1	12	0	36
		2	3	17	3	13	36
		3	8	0	39	3	50
		4	1	6	4	39	50
	%	1	63,9	2,8	33,3	,0	100,0
		2	8,3	47,2	8,3	36,1	100,0
		3	16,0	,0	78,0	6,0	100,0
		4	2,0	12,0	8,0	78,0	100,0

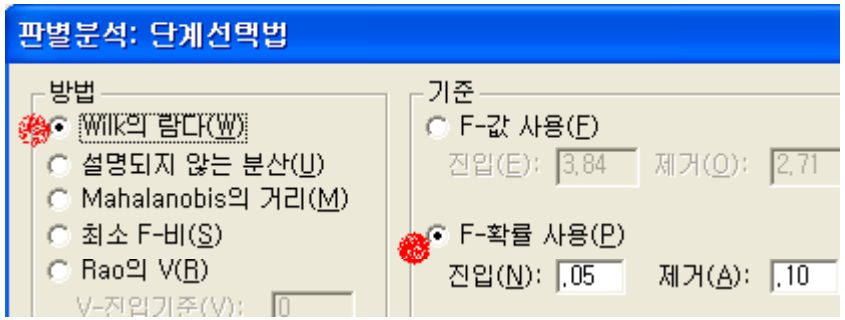
a. 원래의 집단 케이스 중 68.6%이(가) 올바로 분류되었습니다.

새로운 개체 분류

TYPE	R1	R2	R3	R4	D1	D2	D3	D4	Dis	Dis	Dis	Dis	Dis_2
4	55.34	231.0	97.00	43.0	53.06	230.0	98.00	41.0	4	0	0	0	1	4
4	59.65	230.0	89.00	53.0	56.25	227.0	98.00	46.0	4	0	0	0	1	4
	50.00	200.0	90.00	50.0	50.00	250.0	100.0	50.0	.	.	.	.	.	2

■판별 변수 단계 선택


판별분석
집단변수(G):
TYPE(1 4)
범위지정(D)...
독립변수(I):
R1
R2
R3
독립변수들 모두 진
단계선택법 사용(U)
확인
판별분석: 범위지정
최소값(I):
1
최대값(A):
4


판별분석: 단계선택법
방법
Wilk의 람다(W)
설명되지 않는 분산(U)
Mahalanobis의 거리(M)
최소 F-비(S)
Rao의 V(R)
V-진입기준(V):
0
기준
F-값 사용(F)
진입(E):
3.84
제거(O):
2.71
F-확률 사용(P)
진입(N):
.05
제거(A):
.10

■선택된 판별변수

R1
R2
D1
D3
D4

분류결과[a]

		TYPE	예측 소속집단 1	2	3	4	전체
원래값	빈도	1	27	1	8	0	36
		2	4	20	2	10	36
		3	12	0	36	2	50
		4	1	17	2	30	50
	%	1	75,0	2,8	22,2	,0	100,0
		2	11,1	55,6	5,6	27,8	100,0
		3	24,0	,0	72,0	4,0	100,0
		4	2,0	34,0	4,0	60,0	100,0

a. 원래의 집단 케이스 중 65.7%이(가) 올바로 분류되었습니다.

- 개체를 판별하는데 유의한 변수들은 R1, R2, D1, D3. D4이다. 8개 변수를 모두 사용했을 때에 비해 오분류가 31%에서 34%로 조금 증가하였다. 그러나 새로운 개체를 분류하는데 5개 변수만 필요하니 경제적이다.
- 판별에 유의한 변수만으로 개체를 분류하는 것이 오분류 비율 면에서도 더 효율적일 경우도 빈번히 발생하므로 두 방법 모두 사용하여 오분류 비율과 오분류 형태를 보고 분석자가 어느 판별 함수를 사용할지 판단하면 된다.

■판별식에 대한 정보 얻기

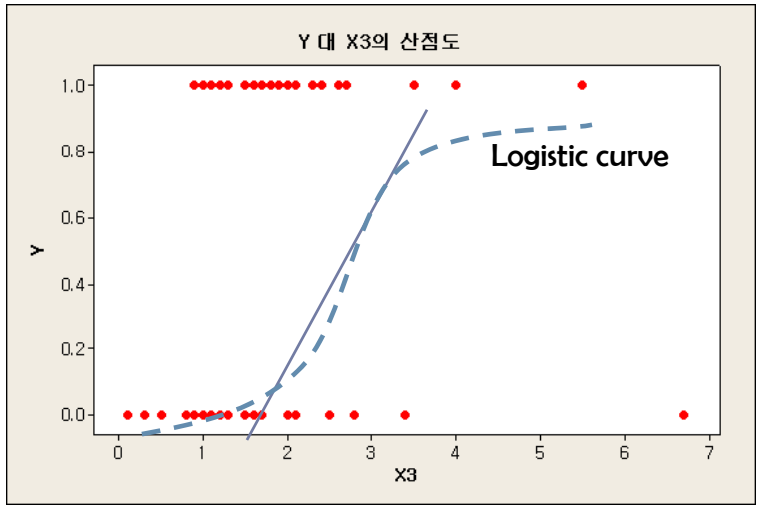
- 반드시 주성분 분석(판별변수가 4개 이상) 혹은 판별변수의 산점도 활용
- 오분류 개체에 대한 정보를 얻는다.

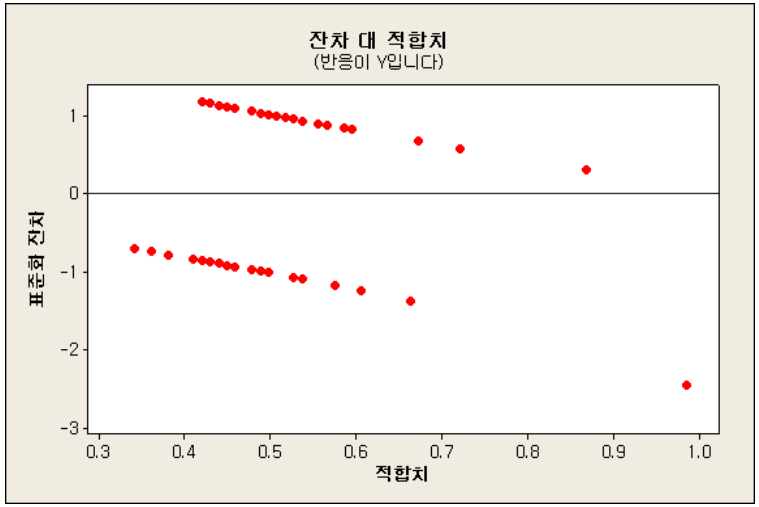
개념

- 종속변수가 metric이 아닌 경우 회귀분석
  - 종속변수 형태: 이진형(binary), 순서형(ordinal), 명목형(nominal)

일반회귀분석 문제점

- 기업의 향후 지불 능력(Y, 0=2년 후 파산, 1=지불능력 있음)에 재무 관련 변수(X1, X2, X3)가 영향을 미칠 것이라 판단
- X1=(보유이익/총자산), X2=(과세전 수익/총자산), X3=(매출액/총자산)
- Y=X3 단순회귀 산점도 및 잔차 산점도(잔차 vs. 예측치) 결과





종속변수 변환 이유

- 종속변수를 metric 변수화, 이론적으로 가질 수 있는 값의 범위 (-∞, ∞)
- p=성공확률 $p_i = \Pr(Y=1 \mid \underline{x})$
- Odds 비율: 도박 배당 기준

$$ODDs = \frac{p}{1-p}$$

변환 방법

- LOGIT 변환 $\ln(\frac{p_i}{1-p_i}) = a + \sum b_k x_{ki} + e_i$
- Probit 변환 $\Phi^{-1}(p_i) = a + \sum b_k x_{ki} + e_i$
- Gompertz 변환 $\ln(-\ln(1-p_i)) = a + \sum b_k x_{ki} + e_i$

$$y_i = \ln(\frac{p_i}{1-p_i}) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + .. + \beta_p x_{pi} + e_i$$

어느 변환 방법 사용?

- Logit 변환 가장 많이 사용
  - 회귀계수의 부호는 성공확률(p) 증감과 일치
  - EXP(회귀계수)는 설명변수가 한 단위 증가할 때 odds ratio에 미치는 영향(multiplication)이 된다.

$$p_i = \frac{1}{1+e^{-\{\alpha+\beta_1 x_{1i}+\beta_2 x_{2i}+..+\beta_p x_{pi}\}}}$$

$$\frac{p_i}{1-p_i} = (e^{\hat{\alpha}})(e^{\hat{\beta}_1})^{x_{1i}} \cdots (e^{\hat{\beta}_p})^{x_{pi}}$$

■판별분석에 활용

- 로지스틱 회귀분석 결과는 추정 회귀식으로부터 각 집단에 소속될 확률이 추정된다.

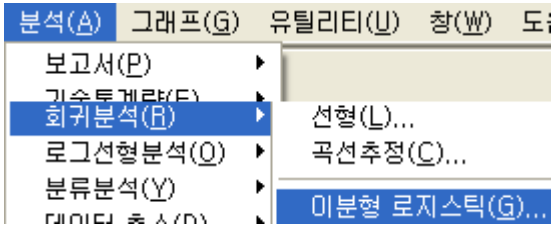
$$p = P(Y = \text{"성공"}) = \frac{1}{1+e^{-\{\alpha+\beta_1 x_{1i}+\beta_2 x_{2i}+..+\beta_p x_{pi}\}}}$$

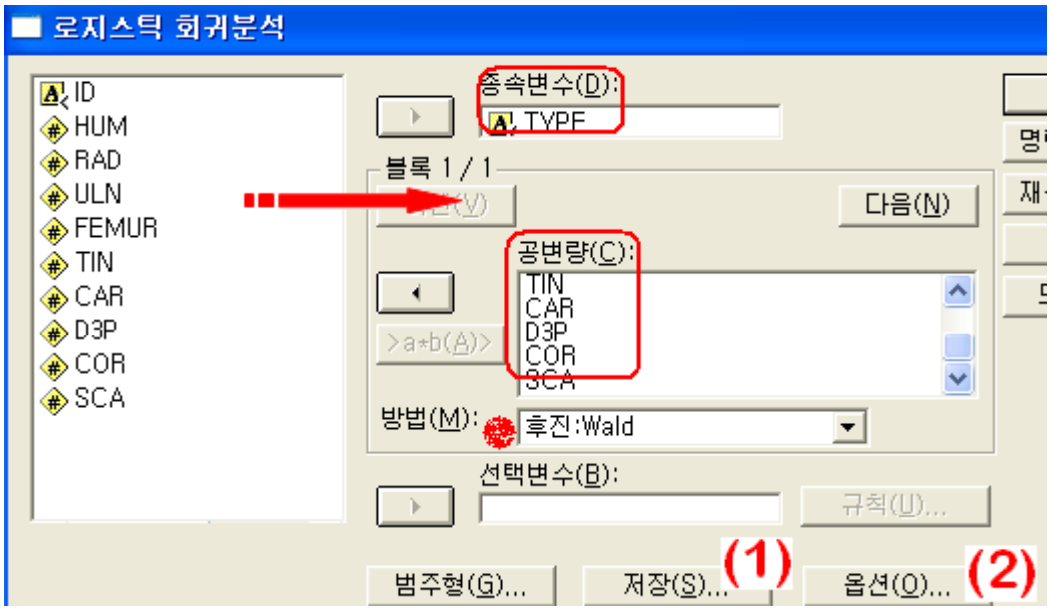
- 판별분석은 판별 변수가 모두 측정형인 경우 사용할 수 있다. 물론 decision tree 방법(CART, CHAID)인 경우 판별 변수가 이산형이나 순서형 분류형 변수인 경우도 가능하지만 일반적으로 측정형 변수만이 판별에 이용된다.

- 로지스틱 회귀 분석(Logistic Regression)은 종속 변수가 이진형(binary, dichotomous: 가질 수 있는 값이 0또는 1인 변수)이거나 순서형(ordinal: 상/중/하) 변수인 경우 사용되는 회귀 분석이다. 그러므로 판별 변수가 설명 변수이고 종속 변수가 집단이 된다. 회귀 분석의 변수 선택 방법에 의해 유의한 판별 변수를 선택하면 되고 판별 변수가 측정형 변수가 아니더라도 판별 변수로 사용할 수 있다.

- 로지스틱 회귀분석은 이진형 반응변수뿐 아니라 반응변수가 순서형(ordinal) 분류형인 경우 사용할 수 있다. 예를 들면 종속 변수가 고객의 신용도이고 이 변수가 (상, 중, 하) 분류되어 있는 경우 사용할 수 있다. SPSS는 이진형 로지스틱 회귀분석만 제공한다.

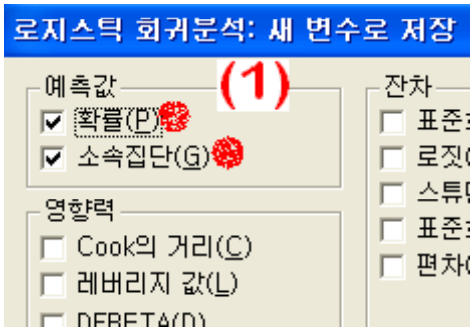
TURKEY.SAV 예제 계속

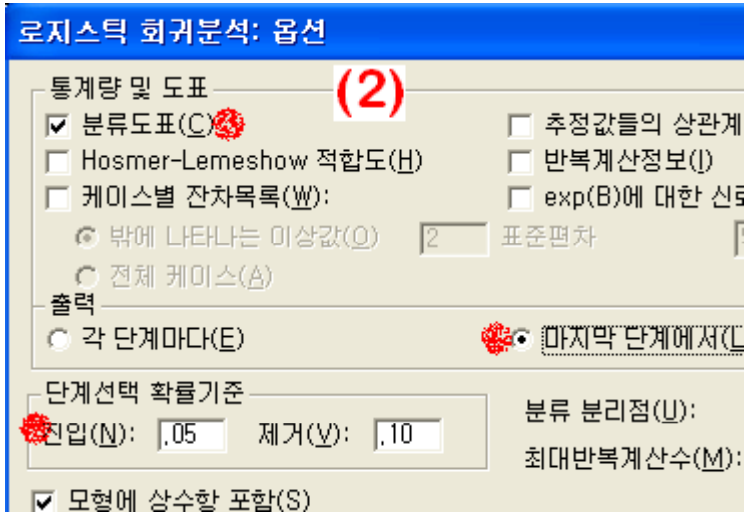




- "1"의 범주형 옵션은 설명변수 중 범주형 변수를 지정한다. 로지스틱 회귀분석에서 범주형 설명변수는 회귀분석과 마찬가지로 가변수(지시변수)처럼 다루어야 한다.
- "2"에서는 성공 확률과 어느 집단으로 판별되는지 데이터에 변수로 저장되도록 설정하였다. "3"에서는 오분류 표와 추정결과 최종 단계만 출력되게 하였다.

▪ 1과 2의 옵션 설정





변수 선택 과정이 출력되었다. 왜 모두 다 출력하는지... 9단계가 최적 단계이다. 10단계 결과를 보면 새로 들어간 FEMUR 변수는 유의하지 않다. B는 표준화 회귀 계수를 의미한다. TIN의 부호가 음이므로 TIN의 크기가 클수록 성공 확률이 높아진다(사육질면조일 가능성이 높다). 오분류 비율은 13%로 Fisher의 판별분석 방법 31%에 비해 줄었다.

방정식에 포함된 변수[c]

		B	S.E.	Wald	자유도	유의확률	Exp(B)
1 단계	HUM	-,757	18954,895	,000	1	1,000	,469
	RAD	-9,196	19379,886	,000	1	1,000	,000
	LIN	6,402	3518,262	,000	1	,999	603,035
(중간생략)							
9 단계	TIN	-1,056	,617	2,935	1	,087	,348
	상수	82,992	46,995	3,119	1	,077	1,104E+36
	TIN	-,477	,179	7,064	1	,008	,621
	상수	69,404	26,129	7,055	1	,008	1,386E+30
10 단계	FEMUR	1,379	1,029	1,797	1	,180	3,971
	TIN	-3,361	2,457	1,872	1	,171	,035
	상수	301,566	222,224	1,842	1	,175	9,298+130

■TIN만 판별변수로 선택

분류표[a]

			예측값		
			TYPE		
관측			WILD	DOMESTIC	분류정확 %
1 단계	TYPE	WILD	14	0	100.0
		DOMESTIC	0	17	100.0
	전체 %				100.0
2 단계	TYPE	WILD	14	0	100.0
		DOMESTIC	0	17	100.0
(중간생략)	전체 %				87.1
9 단계	TYPE	WILD	12	2	85.7
		DOMESTIC	2	15	88.2
	전체 %				87.1
10 단계	TYPE	WILD	13	1	92.9
		DOMESTIC	0	17	100.0
	전체 %				96.8

■유의수준을 0.15로 하였을 때

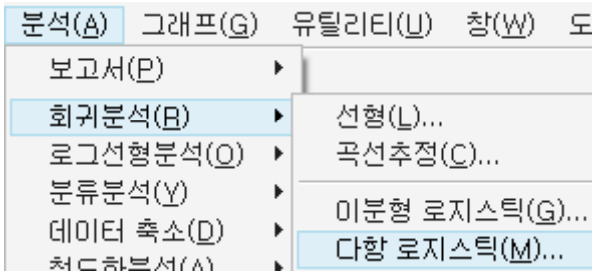
7 단계	ULN	1.157	.758	2.327	1	.127	3.180
	TIN	-1.180	.758	2.420	1	.120	.307
	COR	-.533	.404	1.738	1	.187	.587
	상수	58.890	47.648	1.528	1	.216	3.765E+25

7 단계	TYPE	WILD	13	1	92.9
		DOMESTIC	1	16	94.1
	전체 %				93.5

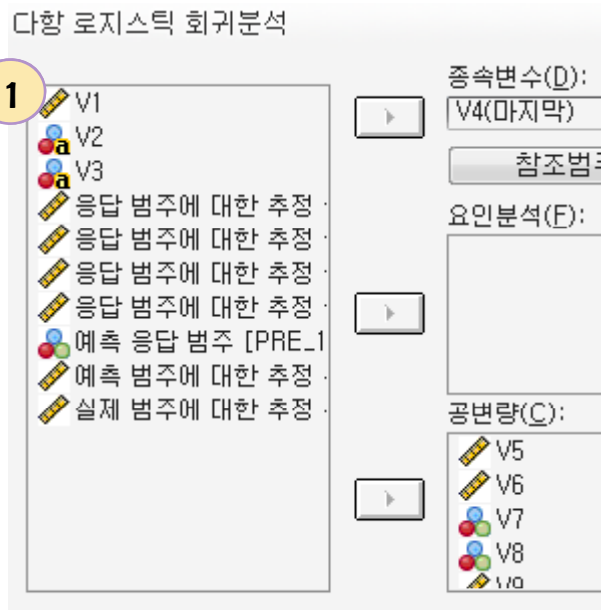
□새로운 개체 판별은 이전과 동일하다. 판별 변수를 ULN, TIN, COR로 하고 새로운 개체에 대한 데이터를 마지막 행에 입력한 후 최종 판별분석을 실시하면 집단이 판별된다. 물론 집단 변수는 결측치로 입력한다.

WHEAT.SAV

- 밀 4 종류

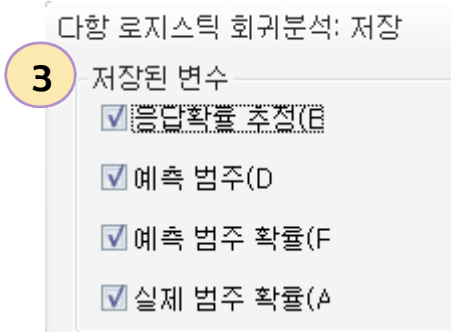


1



2

3



■판별변수 유의성

우도비 검정

효과	모델 맞춤 기준	우도비 검정		
	축소모형의 -2 Log 우도	카이제곱	자유도	유의확률
절편	294,294	38,294	3	,000
V5	332,226	76,226	3	,000
V6	266,569	10,569	3	,014
V7	258,705	2,705	3	,439
V8	261,829	5,829	3	,120
V9	297,694	41,694	3	,000
V10	258,328	2,328	3	,507
V11	268,356	12,357	3	,006
V12	270,350	14,350	3	,002

▸ 유의하지 않은 판별변수 X10, X7, X8 제외

우도비 검정

효과	모델 맞춤 기준	우도비 검정		
	축소모형의 -2 Log 우도	카이제곱	자유도	유의확률
절편	305,282	40,581	3	,000
V5	360,269	95,567	3	,000
V6	277,152	12,450	3	,006
V9	316,121	51,420	3	,000
V11	283,801	19,099	3	,000
V12	278,110	13,408	3	,004

분류

관측수준	예측수준				
	1	2	3	4	분류정확 %
1	24	1	10	1	66,7%
2	4	18	1	13	50,0%
3	7	1	40	2	80,0%
4	2	7	3	38	76,0%
전체 %	21,5%	15,7%	31,4%	31,4%	69,8%

Fisher의 판별분석 방법 66%(슬라이드 54)에 비해 조금 향상

집단 소속 확률

EST1_3	EST2_3	EST3_3	EST4_3	PRE_3	PCP_3	ACP_3
.46	.01	.52	.01	3	.52	.46
.29	.03	.62	.06	3	.62	.29
.33	.01	.65	.01	3	.65	.33
.61	.09	.25	.05	1	.61	.61
.92	.03	.05	.00	1	.92	.92
.79	.15	.03	.02	1	.79	.79
.90	.00	.09	.00	1	.90	.90
.34	.22	.30	.13	1	.34	.34
.69	.01	.29	.00	1	.69	.69

- PCP는 예측집단 속할 확률, ACP는 원 집단에 속할 확률

오분류 개체 진단

- 오분류 개체 표현
  - 집단(정분류, 오분류) 핀별 변수별 평균 비교
  - 판별변수 산점도 (ID, 집단)
  - 주성분 변수 산점도 (ID, 집단)

정준판별분석 (canonical discriminant analysis)

- Fisher에 의해 제안된 방법으로 Fisher's between-within method라고 불리는 방법
- 판별 변수들의 유용한 정보를 모두를 포함한 정준 (Canonical) 변수를 이용하여 판별 분석을 실시한다.
- 판별 변수들의 수가( ) 너무 많아 판별 결과에 대한 해석이 곤란한 경우 -차원 공간에서의 개체들의 집단 평균들을 저 차원 공간으로 변환시켜 처리하는 판별 분석 방법이다.
- 개체 분류가 목적이 아니라 개체 분류 해석을 위해 저 차원으로 표현 ▷주성분분석(개체에 대한 2차원 표현)과 유사

K Nearest Neighbor 판별 분석

- 분류하려는 개체와 Mahalanobis 거리가 가장 가까운 개체 k개를 구하고 그 개체가 속한 집단으로 분류한다.

Classification Trees 방법

- Breiman, Friedman, Olshen, Stone (1984) 제안한 방법인 CART(Classification And Regression Trees)
- J. A. Hartigan가 제안한 CHAID(Chi-square Automatic Interaction Detector)
- Data Mining의 판별분석 기법으로 활용: SAS E/Minor, SPSS Clementine

Individual Directed Technique

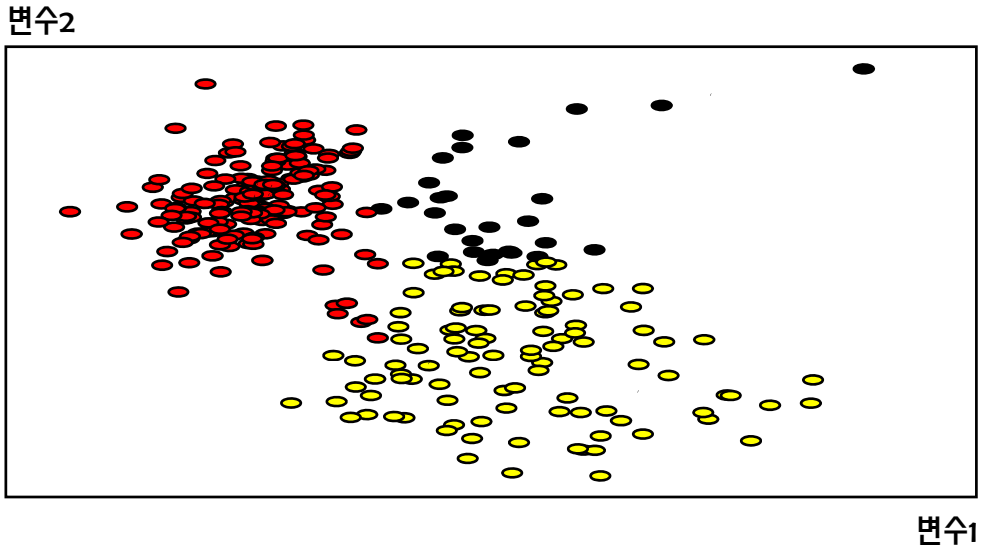
- 범주(그룹)에 대한 사전 정보가 없음
- 다변량 측정치를 동시에 고려하여 데이터 개체 분류
  - 개체의 유사성(similarity, 거리의 반대 개념)을 측정변수들을 이용하여 계산
  - 유사성이 높은 개체를 군집으로 묶어간다.
- 개체를 집단으로 그룹화 하여 각 집단의 성격을 파악함으로써 데이터 전체의 구조에 대한 이해를 얻는 분석 기법
  - 동일 군집 내의 관찰치는 서로 비슷한 속성을 갖도록 하고 서로 다른 군집에 속한 관찰치는 상이한 속성을 갖도록 군집을 구성

군집 원칙

- 동일 군집에 속한 개체 유사한 속성 많음
- 다른 군집에 속하면 유사성 매우 낮음

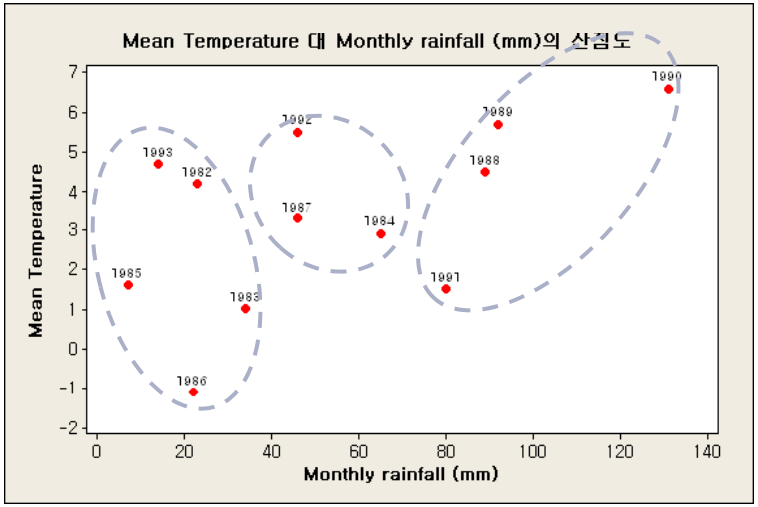
데이터 유형

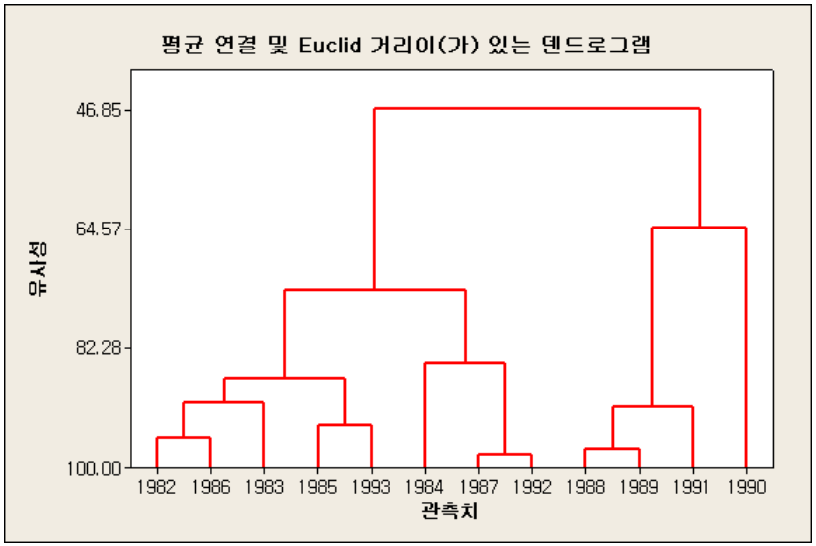
- 측정변수: 측정형(등간 척도 포함)
  - 개체의 속성을 판단하는 기준



CA 목적

- 유사한 성향을 가진 개체를 모아 군집을 형성
- 시각적 표현(주성분 분석 이용)을 통하여 군집간의 특성을 관찰하거나 목표변수와 관계를 파악
  - 개체를 동질적 속성에 의해 묶음으로써 데이터의 구조를 파악할 수 있음
  - 데이터의 차원을 축약하여 이용할 수 있음
  - 개체를 분류하기 위한 명확한 분류기준이 존재하지 않거나 기준이 밝혀지지 않은 상태에서 유용하게 이용





장점

- 탐색적인 기법: 주어진 자료의 내부구조에 대한 사전정보 없이 의미 있는 자료구조를 찾아낼 수 있음
- 다양한 형태의 데이터에 적용가능: 유사성(거리)만 정의되면 모든 종류(텍스트 데이터)의 자료에 적용할 수 있음.
- 분석방법 적용 용이성: 자료의 사전정보를 필요로 하지 않아서 누구나 쉽게 분석할 수 있음

단점

- 가중치와 거리 정의: 가중치와 거리를 어떻게 정의하는 가에 따라 군집분석의 결과가 아주 민감하게 반응함
- 초기 군집 수의 결정이나, 군집 개수 결정이 쉽지 않음
- 결과의 해석이 어려움: 찾아진 군집이 무엇을 의미 하는지 데이터만을 이용해서는 알 수가 없는 경우가 많음
  - 주성분 분석을 이용하여 개체 집단 표현
  - 인구학적 특성에 의해 개체 특성 파악

Hierarchical 계층적 방법

- 유사성이 가까운 순서대로 개체들을 묶어(군집화) 가는 방법
- 한계점
  - 한 대상이 일단 어느 군집에 소속되면 다른 군집으로 이동될 수 없음
  - 이상 개체(outlier)는 제거되지 않고 반드시 어느 군집에 속하게 됨

Non-hierarchical 비계층적 방법

- 군집의 중심이 되는 seed 점들 집합을 선택하여 그 seed 점과 유사성이 높은(거리가 가까운) 개체들을 그룹화 방법
- 문제점
  - 사전에 군집(그룹) 수에 대한 예상이 필요하다.
  - 개체 분류는 처음 선정한 seed 점들에 의해 영향을 많이 받아 분석에 따라 분류가 다를 가능성이 있다.
  - 군집의 수와 seed 값의 위치의 결합 조건이 너무 많아 계산이 분류를 위한 계산이 용이하지 않다.

- 계층적 방법에 의해 군집화 하여, 적절한 군집 수와 이상 개체를 결정한다.
- 이상 개체 제외하고 결정된 군집 수를 이용하여 비계층적 방법에 의해 군집화 한다.

판별분석 차이

- 판별분석은 사전 집단 정보가 있는 경우 집단들간의 차별적 특성을 설명하는 변수들을 발견하여 판별식 유도
- 군집분석은 사전에 집단이 나누어져 있지 않으며 변수를 이용하여 개체 유사성 측정하고 개체를 집단화

요인분석 차이

- 데이터의 구조를 평가한다는 점에서 요인분석에 비유될 수 있으나
- 요인분석은 변수 그룹화, 군집분석은 개체의 그룹화
  - 데이터를 전치하여 변수의 유사성으로 변수 분류 가능

←변수 (p개): 개체 속성→

↑ 개체 (n개) ↓

$$\begin{bmatrix} 1 \\ 2 \\ \cdots \\ n \end{bmatrix} \begin{pmatrix} X_{11} X_{12} \cdots X_{1p} \\ X_{21} X_{22} \cdots X_{2p} \\ \cdots \\ X_{n1} X_{n2} \cdots X_{np} \end{pmatrix}$$

유의사항

- 체계적인 통계적 추론에 의해 개발되지 않아 상대적으로 단순한 절차로 그 결과가 검정되지 못한 경우가 많음
- 동일한 표본에 대하여 상이한 군집분석 알고리즘을 사용하는 경우 상이한 결과가 만들어 질 수 있음
- 많은 경우에 있어서 거리측정방법이 달라지면 군집분석의 결과도 달라짐. 가능하면 몇 가지 측정방법을 사용하여서 이 결과를 이론적인 내용이나 기존의 연구결과와 비교해서 평가하는 것이 바람직함
- 변수간의 측정 척도가 상이한 경우에는 군집분석을 행하기 전에 표준화 하는 것이 바람직함. 이러한 표준화는 특정 변수의 변화 정도가 다른 변수에 비해 특히 큰 경우에 바람직함

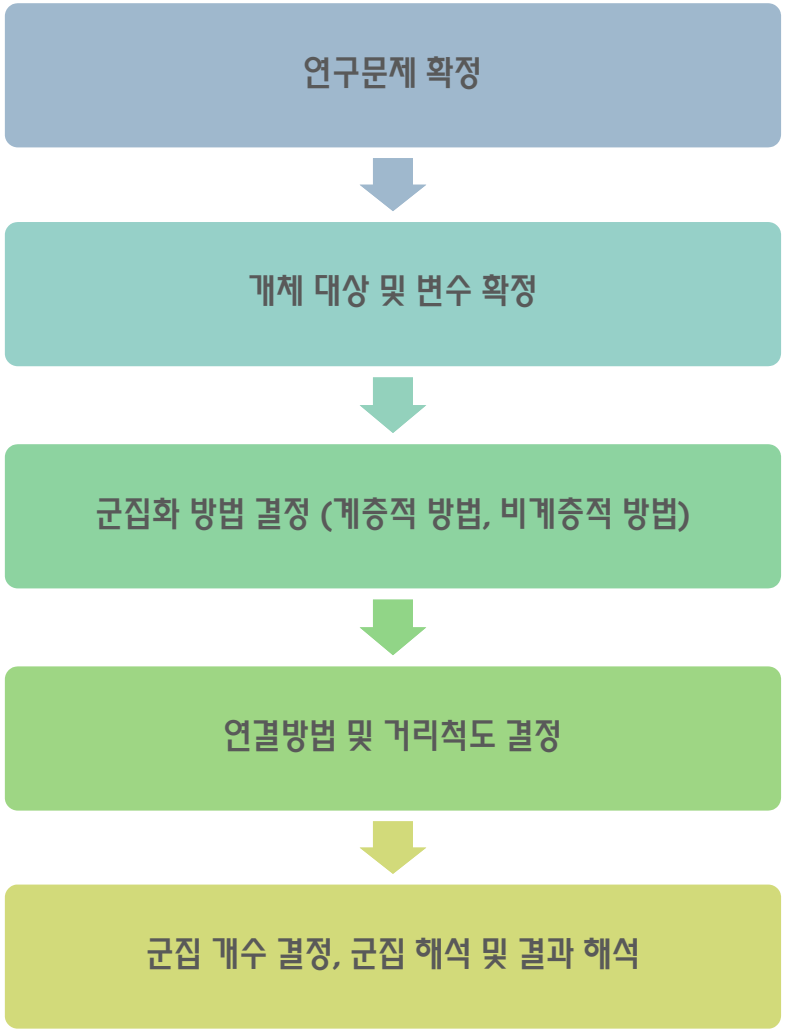
■휴대폰 신기능의 수요층에 대한 고객 세분화 사례

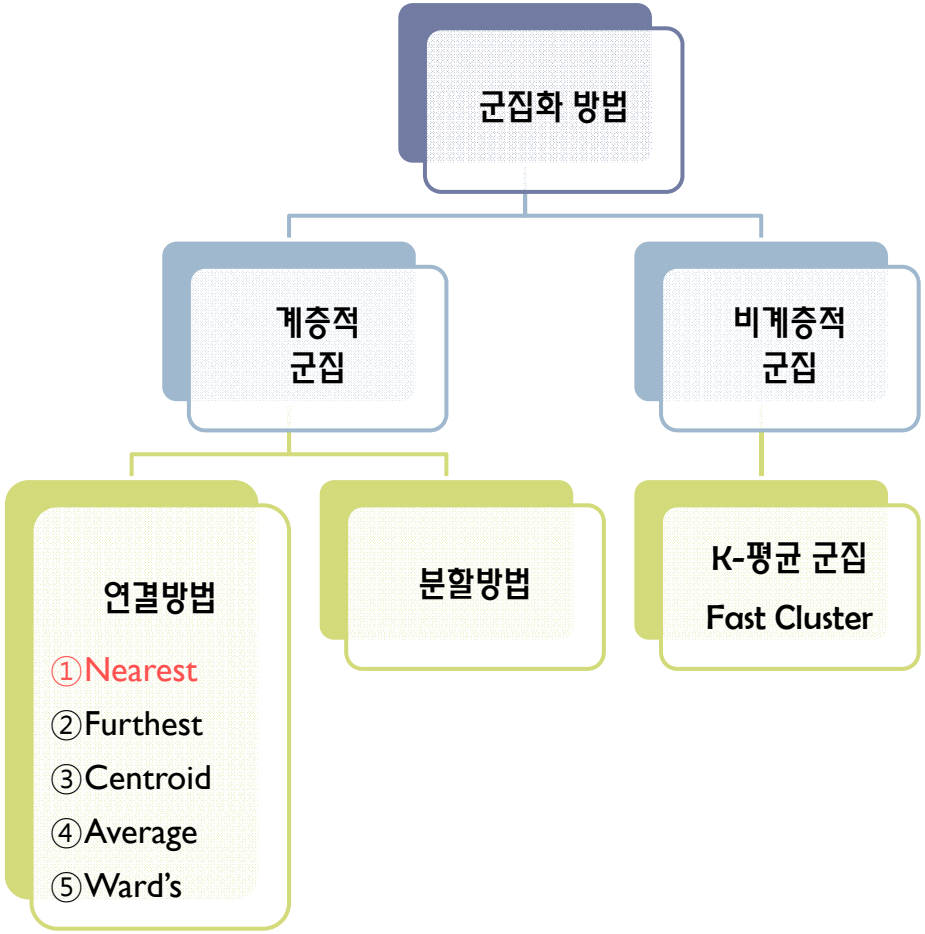
	고객 행동	인구학적 변인
항목	X1: 신기술이 시장에 나올 때마다 나는 최초로 그 기술을 채택한다. (1: 전혀 아니다 ~ 7:매우 그렇다) X2: 카메라 기능이 휴대폰 구매에 어느 정도 영향을 미치십니까? (1: 전혀 중요하지 않다. ~ 7:매우 중요하다) X3: 동영상 기능이 휴대폰 구매에 어느 정도 영향을 미치십니까? (1: 전혀 중요하지 않다. ~ 7:매우 중요하다) X4: 무선인터넷 기능이 휴대폰 구매에 어느 정도 영향을 미치십니까? (1: 전혀 중요하지 않다. ~ 7:매우 중요하다)	Z0: 휴대폰 교체의사 (No:0, Yes:1) Z1: 귀하의 성별은? (남자:1, 여자: 0) Z2: 귀하의 나이는? Z3: 귀하의 주거지역은? (1: 서울/수도권, 2: 광역시, 3: 시/군, 4:기타) 귀하의 직업은? Z4: 중고등학생 (No:0, Yes:1) Z5: 대학생 (No:0, Yes:1) Z6: 전문직 (No:0, Yes:1) Z7: 생산직 (No:0, Yes:1) Z8: 영업직 (No:0, Yes:1)

- 고객의 인구학적 특성(성별, 나이, 직업)으로 군집화하기 보다는
- 고객의 행동패턴(예: 인터넷 사용시간, 한달 통화료 등)을 이용하여 군집화 (군집 이름 부여) 하고
- 군집의 인구학적 특성을 파악함

변량	전체	추종자 CL1	Early Adopter CL2	기본 기능족 CL3
나이	28.6	27.2	19.3	53.0
교체의사	.429	.333	.833	.000
고등학생	.238	.0833	.667	.000
영업직	.238	.333	.000	.333
성별	.619	.750	.333	.667
대학생	.143	.167	.167	.000
전문직	.286	.333	.167	.333
거주지	1.33	1.50	1.17	1.00
생산직	.0476	.0833	.000	.000

추종자	- 20대 후반~30대 초반 - 휴대폰 교체의사는 유보적 - 대학생과 전문직이 70% 차지 - 남자가 75% 점유
Early Adopter	- 10대 후반~20대 초반 - 휴대폰 교체의사는 높음 - 중고생이 대부분 - 여성이 65%를 점유
기본 기능 선호자	- 40~50대 - 휴대폰 교체의사는 거의 없음 - 영업적과 전문직이 66% 점유 - 남성이 65%를 점유


연구문제 확정
개체 대상 및 변수 확정
군집화 방법 결정 (계층적 방법, 비계층적 방법)
연결방법 및 거리척도 결정
군집 개수 결정, 군집 해석 및 결과 해석


군집화 방법
계층적 군집
비계층적 군집
연결방법
①Nearest
②Furthest
③Centroid
④Average
⑤Ward's
분할방법
K-평균 군집
Fast Cluster

개념

- 데이터를 사용하여 유사성이 가장 큰 개체끼리 순차적으로 개체를 분류
- 계층 군집분석의 결과인 덴드로그램 (Dendrogram)을 통해 개체 군집 현황과 전체 군집들간의 구조적 관계 파악
- 군집 이름 부여, 군집 특성 파악: 주성분 분석 활용

주요 원리

- 개체(집단)끼리 유사성(similarity) 측정하여 가장 유사한 개체(혹은 집단)끼리 순차적으로 묶음
  - 전체 대상을 하나의 군집으로 해서 출발하여 개체들을 분할 해 나가는 방법: 분할 (Division) 방법
- 개체간 유사성 정도를 측정하는 개념 필요: 유사성을 거리로 정의
- 집단과 개체(개체) 유사성 정의 필요: 연결(linkage) 방법

유사성 개념

- 데이터 내 속성(변수)면에서 개체의 유사 정도를 나타냄
- 군집분석에서는 비유사성 척도인 거리(distance)를 이용

←변수 (p개): 개체 속성→

↑ 개체 (n개) ↓

$$\begin{bmatrix} 1 \\ 2 \\ \cdots \\ n \end{bmatrix} \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ & & \cdots & \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}$$

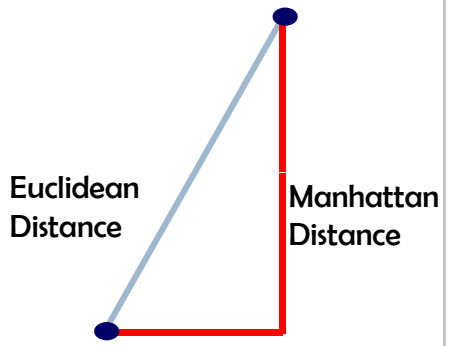
거리의 종류

- 개체 i, 개체 k, j=1,2, ..., p: 군집 변수
- 클리드(Euclidian) (제곱 square)거리
  - 최단 거리, 가장 많이 사용
- 맨하탄(Manhattan) (제곱 square) 거리
  - 직선 이동 거리, 이상치 비중 약해짐
- 피어슨(Pearson) 거리
  - 거리를 변수 분산으로 나누어 표준화 개념

$$d(i,k) = \sqrt{\sum_j (x_{ij} - x_{kj})^2}$$

$$d(i,k) = \sum_j |x_{ij} - x_{kj}|$$

$$d(i,k) = \sqrt{\sum_j (x_{ij} - x_{kj})^2 / v_j}$$



변수 표준화

- 군집 변수의 단위가 다르면(분산의 크기 다름) 단위 큰 변량이 개체 거리(유사성)에 영향을 준다.
- 그러므로 변량 단위 통일을 위한 변량 표준화 필요
- Pearson 거리는 표준화 개념이 고려됨

$$Z_{ij} = \frac{X_{ij} - \bar{X}_{.j}}{S(X_{.j})}$$

Linkage

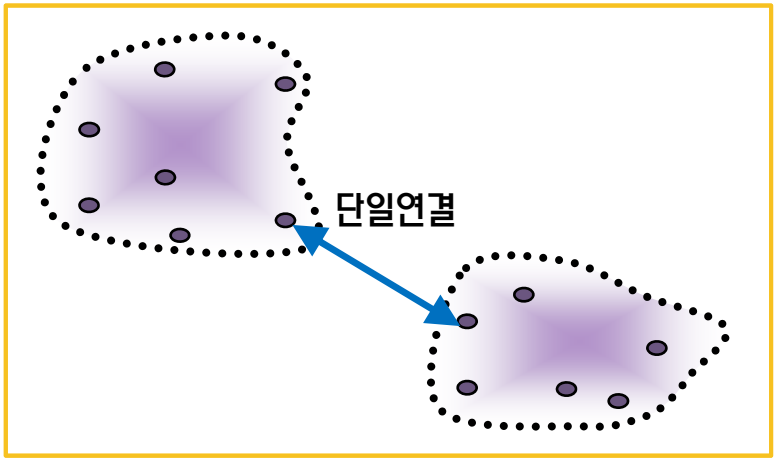
- 가까운 것끼리 순차적으로 묶어갈 때 집단과 개체 (혹은 집단) 거리 측정을 위한 개념

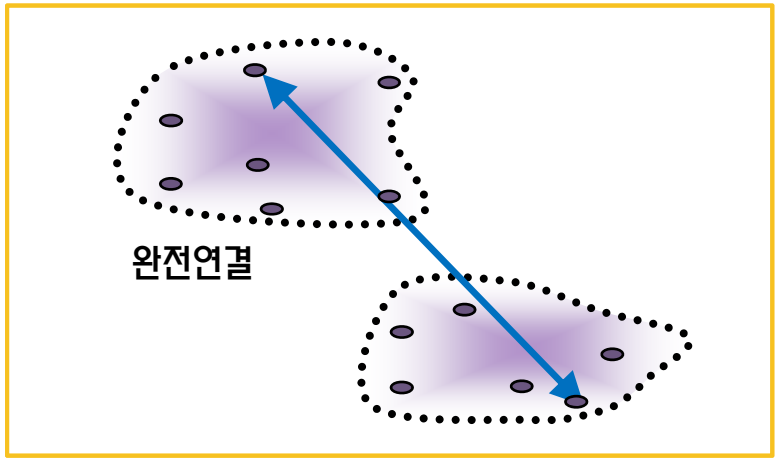
거리 측정 방법

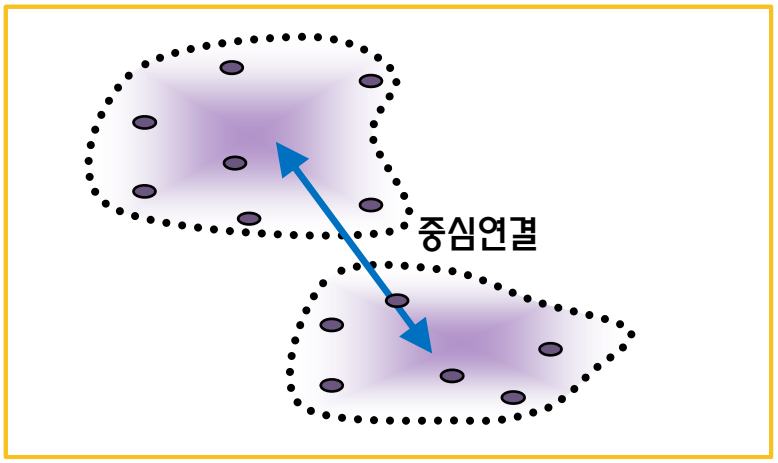
- Nearest neighbor (단일 연결 single): 두 군집의 각 개체 중 가장 가까이 있는 개체의 거리
- Furthest neighbor (완전연결 complete)r: 두 군집의 각 개체 중 가장 멀리 있는 개체의 거리
- Centroid neighbor (중심연결): 군집의 평균 간의 거리
- Average neighbor (평균연결): 한 군집의 개체와 다른 군집 개체들의 각 거리 평균
- Median neighbor (중위수 연결): 평균 대신 거리 중위수 사용, 이상치의 영향 적음
- Ward's minimum variance: 군집의 평균간 거리를 각 군집의 개체 개수의 역의 합으로 나눈 제곱근을 구한 거리

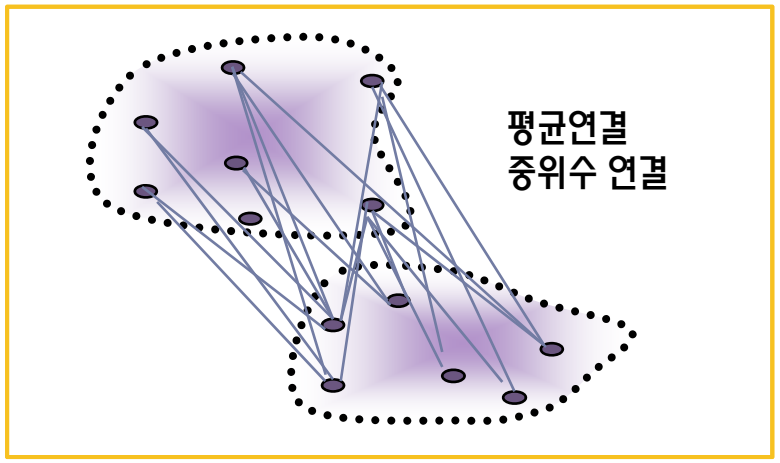
어떤 방법을 사용하는 것이 좋은가?

- Nearest 방법은 군집의 수가 줄어들고 이상 개체 판단에 유리
- Furthest는 군집간 거리를 최소화 하는 경향이 있어 개체 수가 적은 군집을 얻음
- 가장 많이 사용하는 방법은 Average neighbor 방법
- 여러 방법 사용하여
  - 군집간 평균 거리, 군집 내 개체간 평균 거리가 작은 군집 방법


단일연결


완전연결


중심연결


평균연결
중위수 연결

■정의

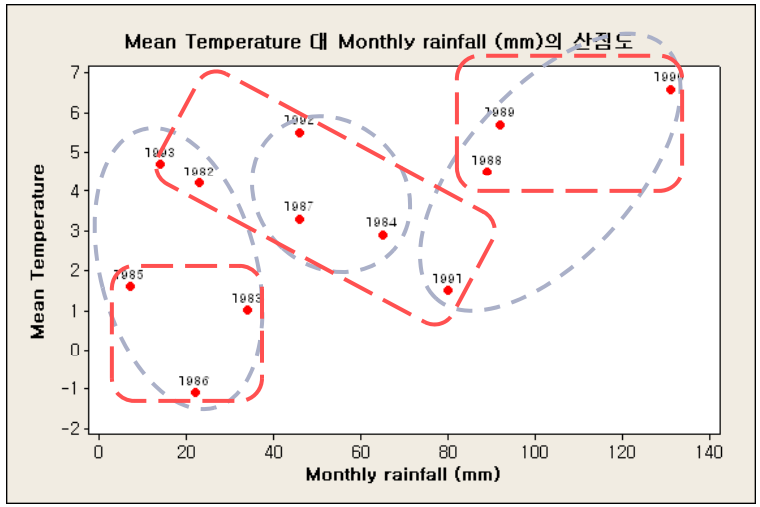
▶군집의 병합 과정 및 집단간 거리를 이차원 도면을 사용하여 간략히 표현

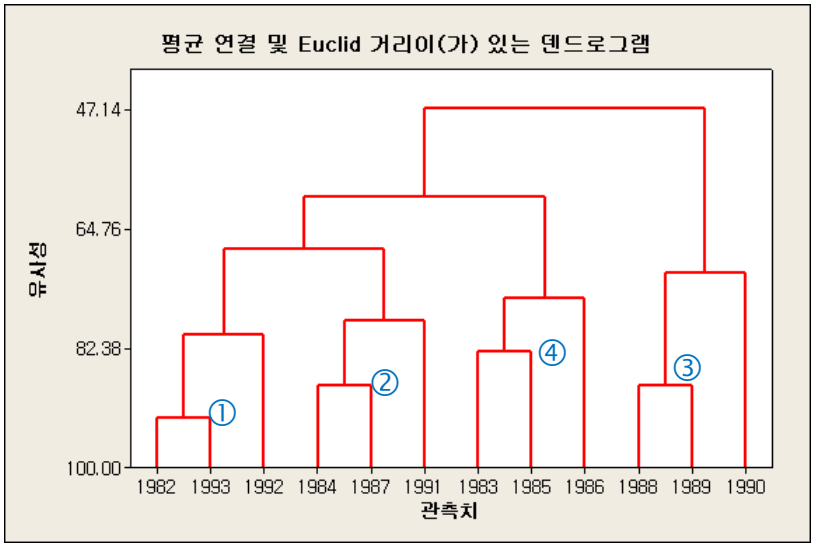
▶유사성이 높은(거리가 가까운) 순서대로 개체를 순차적 연결

▶덴드로그램에서 선의 높이는 유사성 크기를 표시

■예제

▶1982~1993년 2월(n=12)을 평균 강수량과 평균 온도 속성으로 군집화 하자. (변수 표준화)





ID	A	B
1	12	24
2	15	20
3	18	17
4	40	44
5	35	24

$$d\{(C_1, C_2)\} = \min(\, d(x, y) \mid x \in C_1,\ y \in C_2\}$$

	1	2	3	4	5
1	0.00				
2	5.00	0.00			
3	9.22	4.24	0.00		
4	34.41	34.66	34.83	0.00	
5	23.00	20.40	18.38	20.62	0.00

$$d(2,3) = \sqrt{(15-20)^2 + (18-17)^2} = 4.24$$

	1	2,3	4	5
1	0.00			
2,3	5.00	0.00		
4	34.41	34.83	0.00	
5	23.00	18.38	20.62	0.00

$$d\{(1),(2,3)\} = \min\{d_{12}, d_{13}\} = d_{12} = 5.00$$
$$d\{(4),(2,3)\} = \min\{d_{24}, d_{34}\} = d_{34} = 34.83$$
$$d\{(5),(2,3)\} = \min\{d_{25}, d_{35}\} = d_{35} = 18.38$$

	1,(2,3)	4	5
1,(2,3)	0.00		
4	34.41	0.00	
5	18.38	20.62	0.00

$$d\{(4),(1,(2,3)\} = \min\{d_{14}, d_{(23)4}\} = d_{14} = 34.41$$
$$d\{(5),(1,(2,3)\} = \min\{d_{15}, d_{(23)5}\} = d_{(23)5} = 18.38$$

	(1,(2,3)),5	4
(1,(2,3)),5	0.00	
4	20.62	0.00

ID	A	B
1	12	24
2	15	20
3	18	17
4	40	44
5	35	24

$$d\{(C_1, C_2)\} = \max(\, d(x, y) \mid x \in C_1,\ y \in C_2\}$$

	1	2	3	4	5
1	0.00				
2	5.00	0.00			
3	9.22	4.24	0.00		
4	34.41	34.66	34.83	0.00	
5	23.00	20.40	18.38	20.62	0.00

$$d(2,3) = \sqrt{(15-20)^2 + (18-17)^2} = 4.24$$

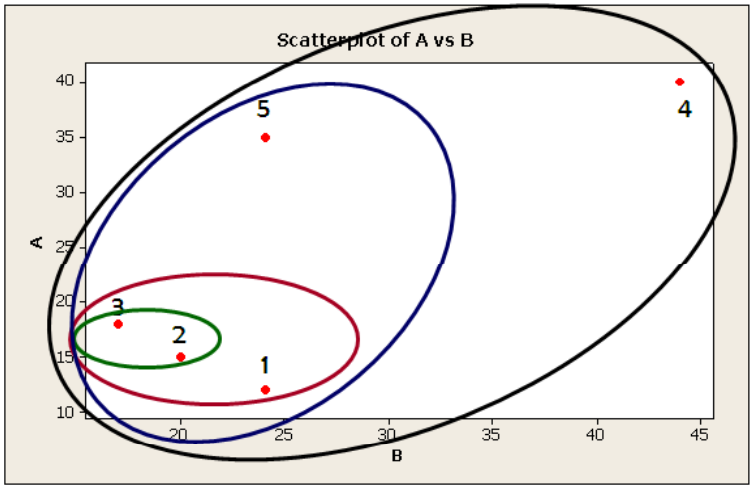
	1	2,3	4	5
1	0.00			
2,3	9.22	0.00		
4	34.41	34.66	0.00	
5	23.00	20.40	20.62	0.00

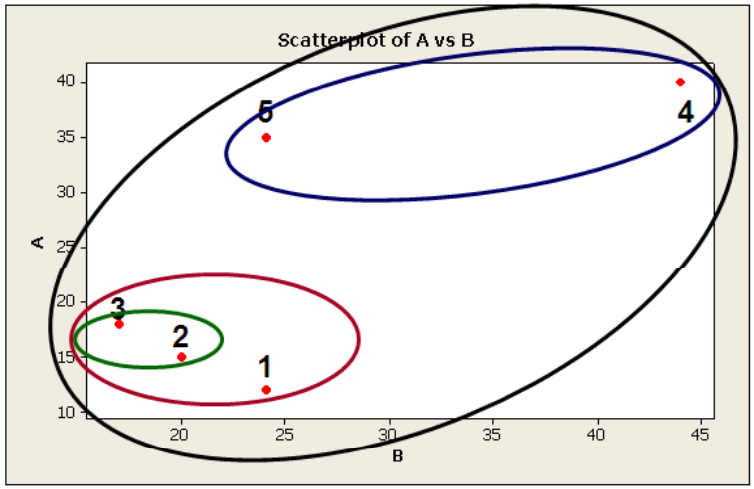
$$d\{(1),(2,3)\} = \max\{d_{12}, d_{13}\} = d_{12} = 9.22$$
$$d\{(4),(2,3)\} = \max\{d_{24}, d_{34}\} = d_{24} = 34.66$$
$$d\{(5),(2,3)\} = \max\{d_{25}, d_{35}\} = d_{25} = 20.40$$

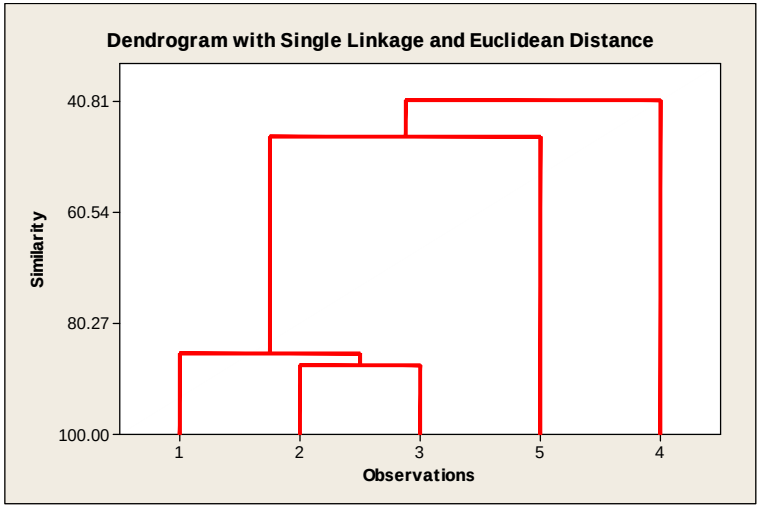
	1,(2,3)	4	5
1,(2,3)	0.00		
4	34.66	0.00	
5	23.00	20.62	0.00

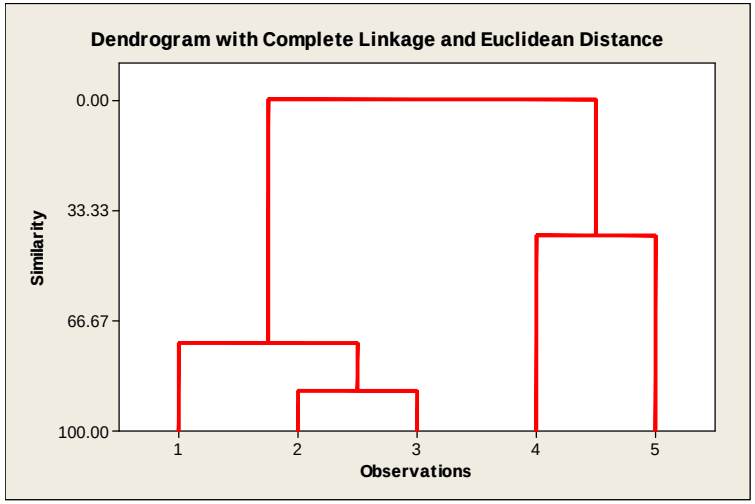
$$d\{(4),(1,(2,3)\} = \max\{d_{14}, d_{(23)4}\} = d_{(23)4} = 34.66$$
$$d\{(5),(1,(2,3)\} = \max\{d_{15}, d_{(23)5}\} = d_{(23)5} = 23.00$$

	(1,(2,3)),5	4
(1,(2,3)),5	0.00	
4	34.83	0.00


Scatterplot of A vs B
A
B
1
2
3
4
5


Scatterplot of A vs B
A
B
1
2
3
4
5

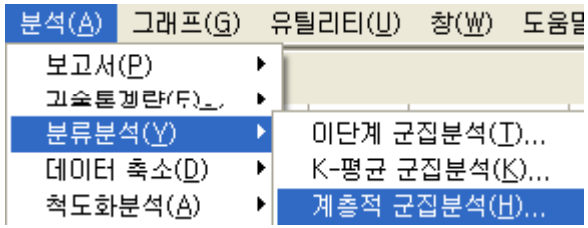

Dendrogram with Single Linkage and Euclidean Distance
Similarity
40.81
60.54
80.27
100.00
1
2
3
5
4
Observations

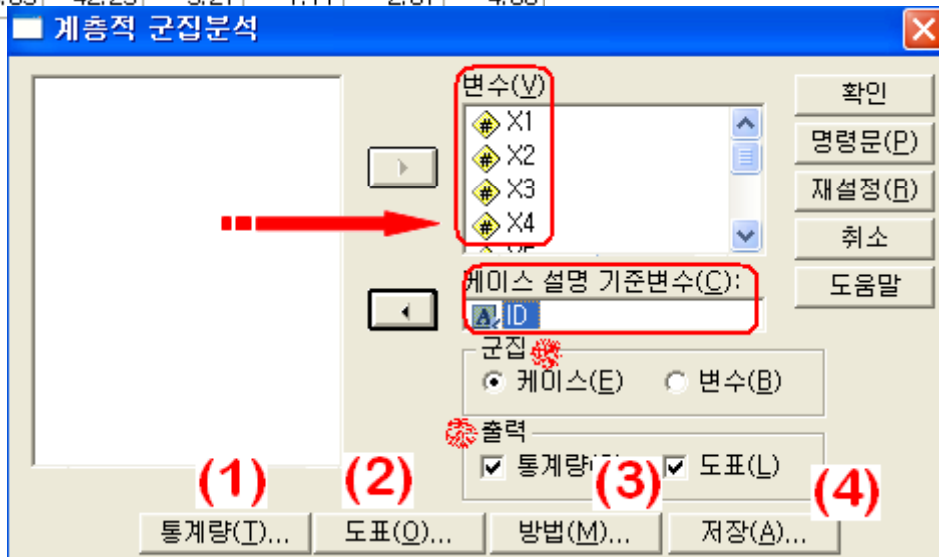

Dendrogram with Complete Linkage and Euclidean Distance
Similarity
0.00
33.33
66.67
100.00
1
2
3
4
5
Observations

PIZZA.SAV

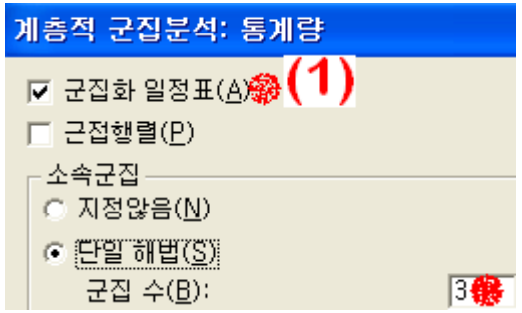
- 56개 피자 제품에 대해 MOIS(수분 함유 X1), PROT(단백질 함유량 X2), FAT(지방 함유량 X3), ASH(ash 함유량 X4), SODIUM(나트륨 함유량 X5), CARB(탄수화물 함유량 X6), CAL(칼로리 X7)를 조사하였다. 이를 이용하여 56개 피자 제품을 분류하여 보자.
- [Applied Multivariate Methods for Data Analysts, Dallas E. Johnson, 1998]

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					





▪ 메뉴 cont.



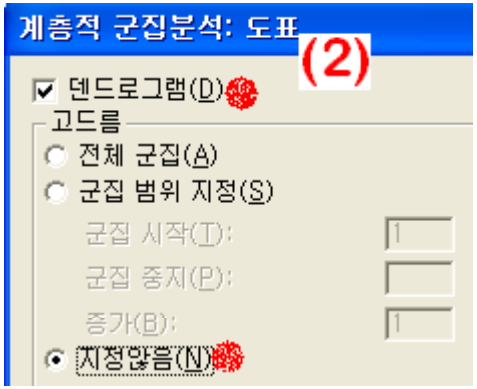
군집화 일정표

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

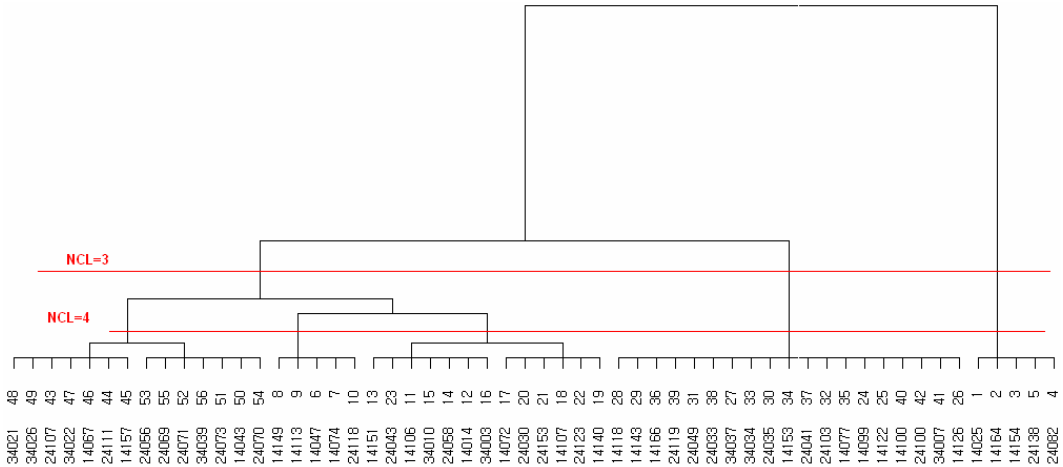
케이스	3 군집
1:14025	1
2:14164	1
3:14154	1
4:24082	1
5:24138	1
6:14047	2
7:14074	2
8:14149	2

□군집화 일정은 가장 가까운 개체부터 묶여지는 과정을 보여준다. 3번째는 군집1과 군집2가 묶였다는 것을 의미한다. 단일 해법에서 군집의 수를 3으로 했으므로 각 개체의 소속 군집이 1, 2, 3 중 하나로 출력된다. "소속 군집"을 굳이 선택할 필요는 없다. Why? 필요하지 않다.

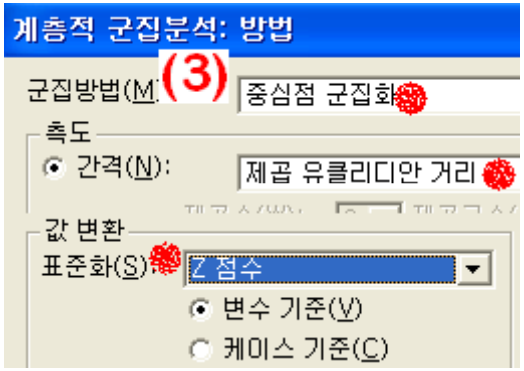


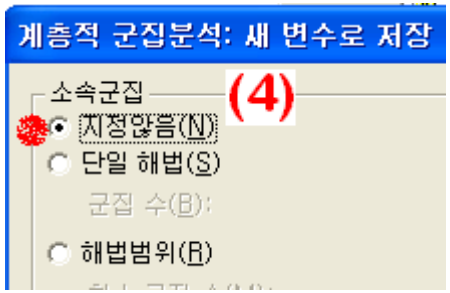
□나무 다이어그램(Dendrogram)을 출력하라는 옵션이다. 고드름 도표는 보기 복잡하고 활용 면에서 떨어지므로 출력하지 않는 것이 좋다. SPSS 출력 창에는 덴드로그램이 수직으로 출력된다. 보기 편하기 하기 위하여 돌려 놓은 것이다.

Dendrogram



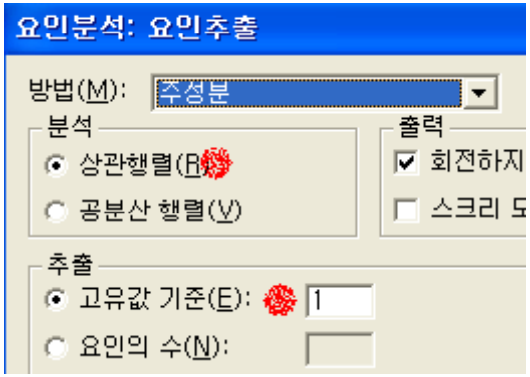
- 군집 방법은 중심점 군집화(Centroid clustering) 방법을 사용하였고 유사성(거리)는 Euclidian 제곱 거리를 사용하였다. 어느 방법이 최선? 아무도 모른다. 값의 표준화 점수를 사용한 이유는 군집에 사용되는 변수의 단위가 다르기 때문이다.
- 덴드로그램을 보고 군집의 개수를 결정하기 전까지는 소속 군집을 출력하지 않는 것이 좋다.

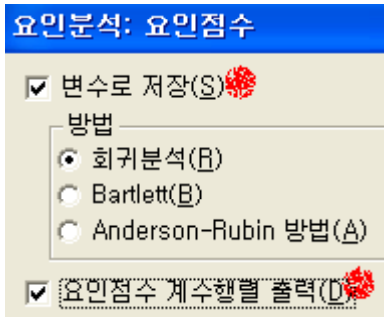




■군집 이름 부여

▪주성분 분석





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

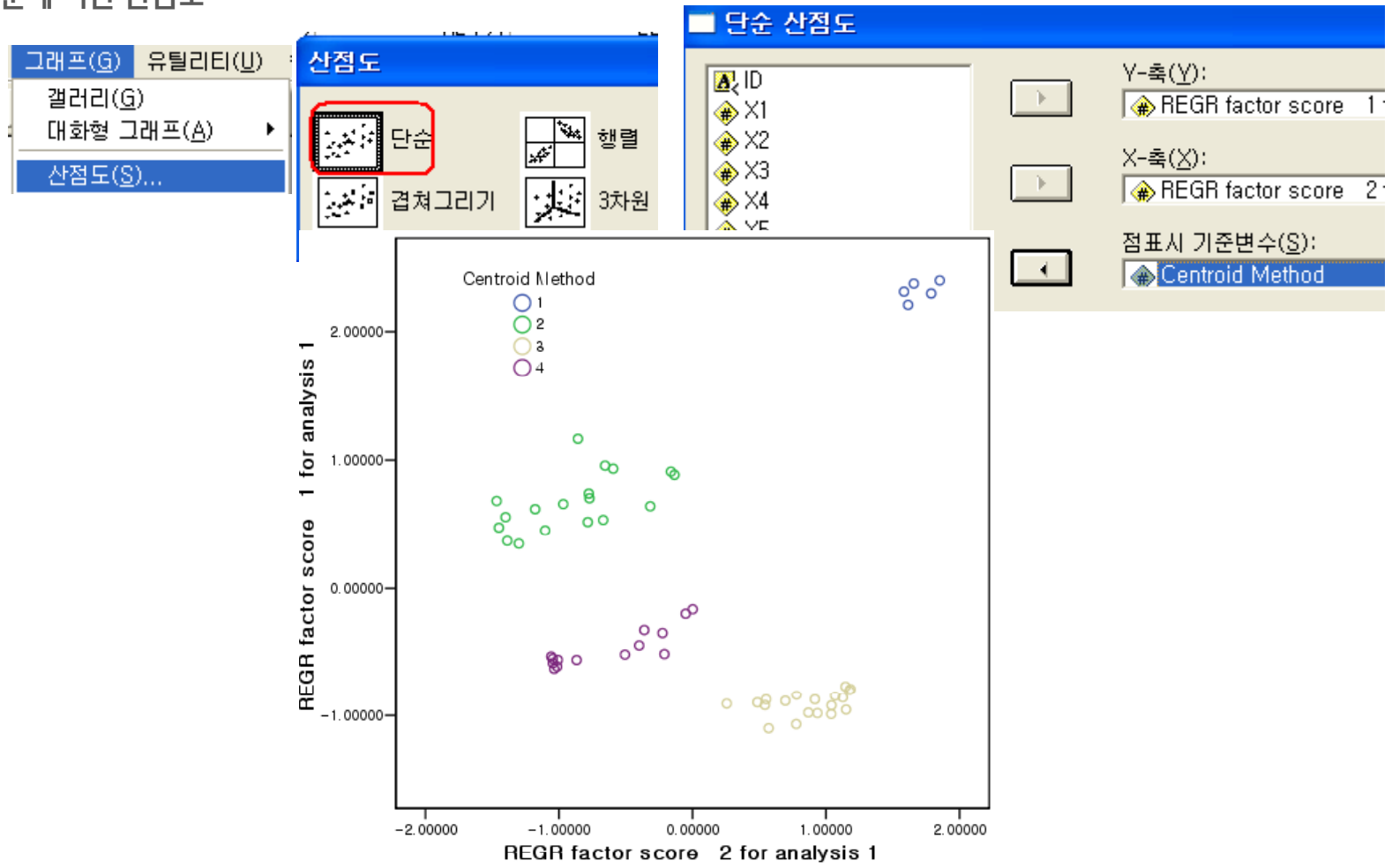
추출 방법: 주성분 분석.

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

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

▪ 주성분에 의한 산점도


그래프(G)
유틸리티(U)
갤러리(G)
대화형 그래프(A)
산점도(S)...
산점도
단순
행렬
겹쳐그리기
3차원
단순 산점도
ID
X1
X2
X3
X4
Y-축(Y):
REGR factor score 1
X-축(X):
REGR factor score 2
점표시 기준변수(S):
Centroid Method
Centroid Method
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

비계층적 군집 정의

- 군집의 개수를 분석 전에 정해야 한다.
  - 계층적 군집, 사전 정보, 분석자의 결정에 의해 군집의 개수 분석 전 결정
- 군집의 중심을 결정
  - 우선 seed(군집의 중심)를 정하고 이 seed에 가까운 개체들을 군집으로 묶는다.
  - 군집이 결합되면, 각 군집별 군집화 과정 오류를 계산한다.
- 군집화 단계에서 오류가 발생하면 seed를 조정하고 오류를 재계산한다.
- 군집화의 각 단계가 끝나면서 발생하는 오류가 발생하지 않으면 군집화를 종료한다.

방법

- 군집의 중심 결정
  - 집단 내 개체 평균: K-means 비계층적 방법
  - Euclidian 거리 중심
- 군집의 크기 결정
  - 반지름(radius) 길이

K-평균 군집 방법

- 사전에 결정된 군집 수 K에 기초하여 각 관측 값을 군집의 중심들 중에서 가장 가까운 군집에 할당하는 방법

단계 1 : 군집의 수 K를 결정

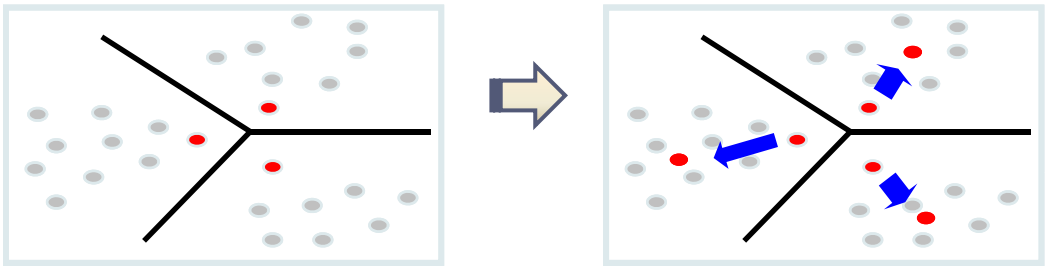
단계 2 : 초기 K개 군집의 중심을 랜덤하게 선택함

단계 3 : 각 관측 값들을 가장 가까운 중심의 군집에 할당함

단계 4 : 새로운 군집에 할당된 관측 값들로 새로운 중심을 계산

단계5: 개체 군집 변동이 없을 때까지 단계 3, 4를 반복한다.

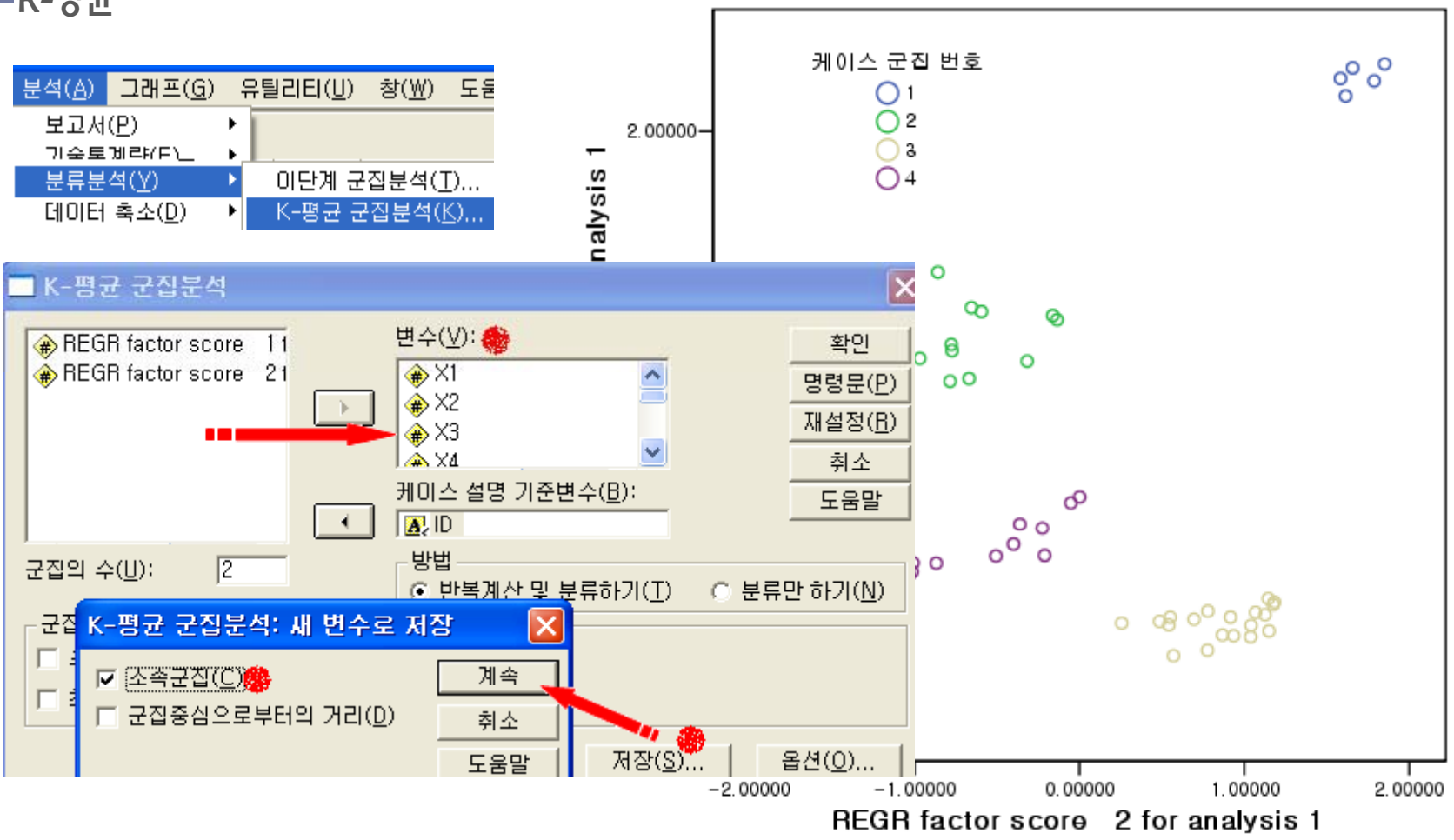
군집화 단계 seed 이동 예 (k=3)



군집 수 K 값 결정

- 계층적 분석 방법에 의해 k 결정
- 여러 k 사용하여 군집간 평균 거리나 군집 내 개체 평균 거리를 활용하여 최적 k 결정

■K-평균


분석(A) 그래프(G) 유틸리티(U) 창(W)
보고서(P)
분류분석(Y)
데이터 축소(D)
이단계 군집분석(T)...
K-평균 군집분석(K)...
K-평균 군집분석
REGR factor score 1
REGR factor score 2
변수(V):
X1
X2
X3
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
-2.00000
-1.00000
0.00000
1.00000
2.00000
REGR factor score 2 for analysis 1

개념

- n개의 개체를 2차원 가시적 공간에 나타내는 방법
- 각 개체간 유사성(similarity) 혹은 거리는 저차원으로 옮겨지더라도 원래 유사성 크기를 갖는다.

유사성 개념

- 개체를 저 차원 가시적 공간(2차원)에 나타내려면 각 개체간 거리(유사성)를 측정해야 한다.
- MDS는 개체(행), 변수(열) 모두 저차원 공간 표현 가능

개체간 유사성 측정

- metric 방법
  - Euclidean distance ► (측정형 변수 거리)
  - 각 개체의 유사성(거리)을 사람들이 리커드 척도나 순위 평가
  - 유사성을 계산하여 개체를 분류하는 면에서는 군집분석과 유사
  - 군집분석은 개체를 군집화 하고, 주성분 점수에 의해 개체를 표현하나, MDS는 유사성에 의해 단지 2차원 공간에 표현
- non-Metric 방법
  - 평가자 들이 개체를 주관적으로 분류하게 하고 그로부터 얻어지는 빈도로부터 유사성을 측정

Metric 방법

- 개체 (i, j) 유사성: 거리 개념 $S_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$
- 변수 유형
  - 개체 속성을 측정하는 변수: 군집분석의 변량과 동일
  - 리커트 척도, 우선 순위: (회사 1, 회사2, 회사3, ...) 혹은 (속성1, 속성2, 속성3, ...)
  - (p≤2)일 때도 유사성(거리)을 계산

개체 \ 변수	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}

Non-metric 방법

- 개체에 대한 빈도표를 이용
- (상대)빈도(f_{ij})가 개체간 유사성 (S_{ij}) 측정

개체 \ 개체	1	2	⋯→	n
1	0			
2	S_{21}	0		
⋮	⋮	⋮	0	
n	S_{n1}	S_{n2}	⋯→	0

- 예제 (non-metric)
  - 유사 상품 분류: 마케팅 전문가 15명에게 n개 상품을 주관적인 판단에 의해 임의로 분류하게 한 후 상대빈도표(?/15) 작성
    - 예: (상품 (1, 2) 유사성 계산) 상품 (1, 2)를 동일 군집으로 분류한 전문가가 10명이라면 유사성은 10/15가 된다.
- 예제 (metric)
  - Under-arm Deodorant생산 회사에서 판매 전략을 세우기 위하여 각 제품들이 서로 얼마나 유사한지(가까운지) 알아보려고 한다. 이를 위하여 소비자를 임의로 선택하여 제품의 각 분야(향기, 냄새 제거 정도, 사용 편리 정도, 옷에 묻어나는 정도)를 10점 만점으로 평가
    - 고객 평가점수에 의한 제품의 유사성 정도를 2차원 공간에 표현
    - 제품 평가 유사성에 따른 고객 세분화

표현 방법

- 각 개체간 유사성(거리)을 측정한다.
- 개체의 개수가 n개인 경우 k=n(n-1)/n개 유사성 그룹이 존재한다.
- 유사성이 작은 것부터 크기 순으로 배열한다. $S_{i1j1} < S_{i2j2} < \ldots < S_{ikjk}$
  - 이를 이용하여 개체를 m(= 2)차원으로 공간으로 줄일 경우 개체간의 거리를 구한다.
- 임의의 한 좌표에 한 개체를 표현하고 나머지 개체들은 상대적 유사성을 고려하여 좌표에 표현한다.

Stress 값

- 2차원 공간으로 줄일 수 있는지를 알아 보는 측정치
- S^r은 차원이 2차원으로 줄었을 때 개체의 유사성

$$stress = \frac{[\sum_{i<j} (S_{ij} - S^{r}_{ij})^2 / S_{ij}]}{\sum_{i<j} (S_{ij})}$$

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

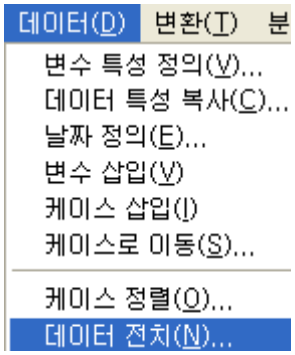
차원(dimension)과 위치(coordination)의 의미

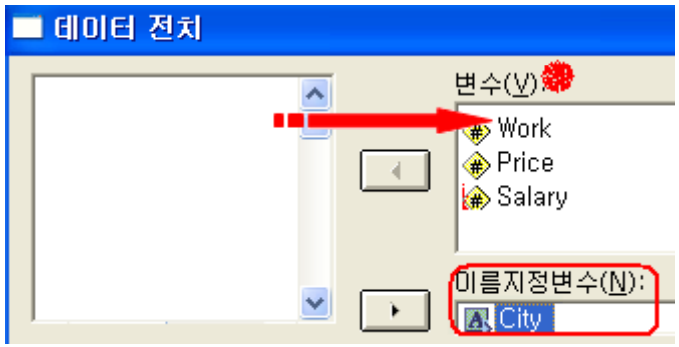
- 아무 의미 없음
- 개체 간 유사성은 얼마나 가까이 있나에 의해 해석됨

CITY.SAV

- 경제적 변인에 의해 도시를 군집화 하려 한다. 도시 이름, 12개 직종 노동시간 가중 평균, 물가, 시간당 임금
- http://lib.stat.cmu.edu/DASL/Datafiles/Cities.html

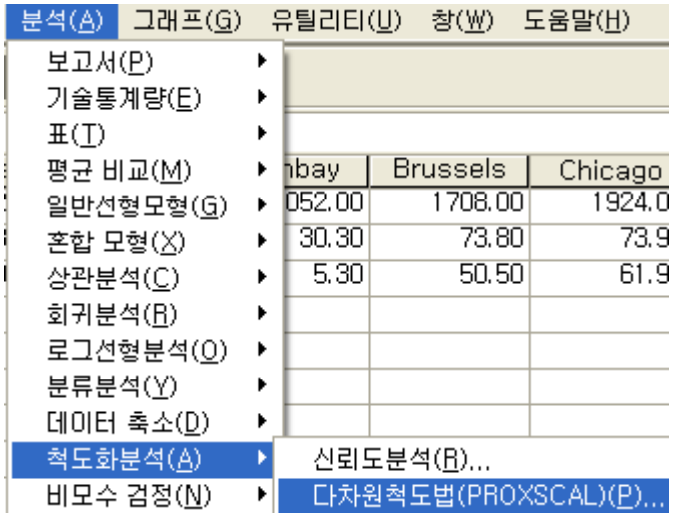
City	Work	Price	Salary
Amsterdam	1714.00	65.60	49.00
Athens	1792.00	53.80	30.40
Bogota	2152.00	37.90	11.50
Bombay	2052.00	30.30	5.30

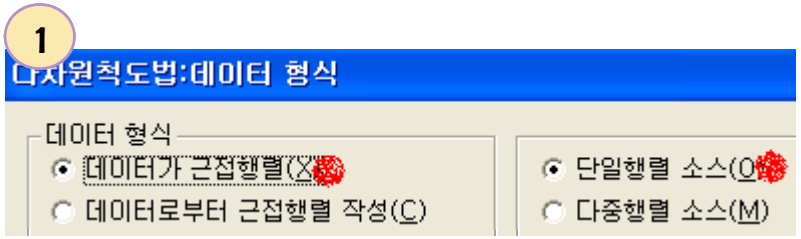




CASE_LBL	Amsterdam	Athens	Bogota	Bombay	Brussels	Chica
Work	1714.00	1792.00	2152.00	2052.00	1708.00	192
Price	65.60	53.80	37.90	30.30	73.80	7
Salary	49.00	30.40	11.50	5.30	50.50	8

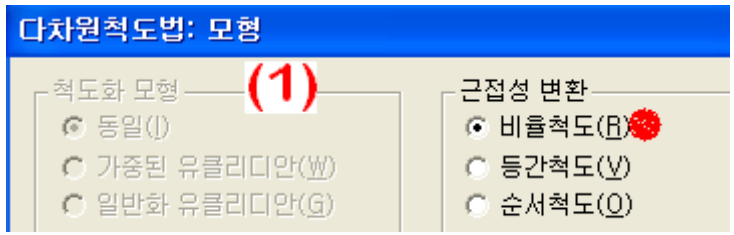
▪ 메뉴


분석(A)
그래프(G)
유틸리티(U)
창(W)
도움말(H)
보고서(P)
기술통계량(E)
표(T)
평균 비교(M)
일반선형모형(G)
혼합 모형(X)
상관분석(C)
회귀분석(R)
로그선형분석(O)
분류분석(Y)
데이터 축소(D)
척도화분석(A)
비모수 검정(N)
신뢰도분석(R)...
다차원척도법(PROXSCAL)(P)...
Brussels
Chicago
1708.00
1924.0
73.80
73.9
50.50
61.9


1
다차원척도법:데이터 형식
데이터 형식
데이터가 근접행렬(X)
데이터로부터 근접행렬 작성(C)
단일행렬 소스(O)
다중행렬 소스(M)


2
변수(V):
Amsterdam
Athens
Bogota
Bombay
Brussels
Chicago
Copenhagen
소스(S):
거리작성에
(1)
측도(A)...
유클리디안 거리
모형(M)...
제한(E)...
옵션(O)...
도표(T)...

메뉴 cont.

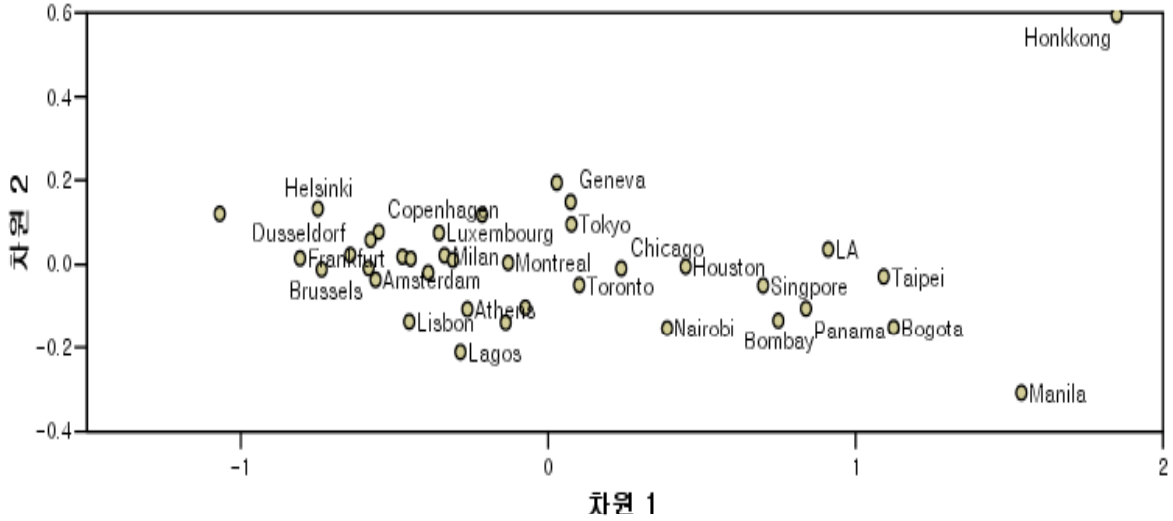


최종좌표

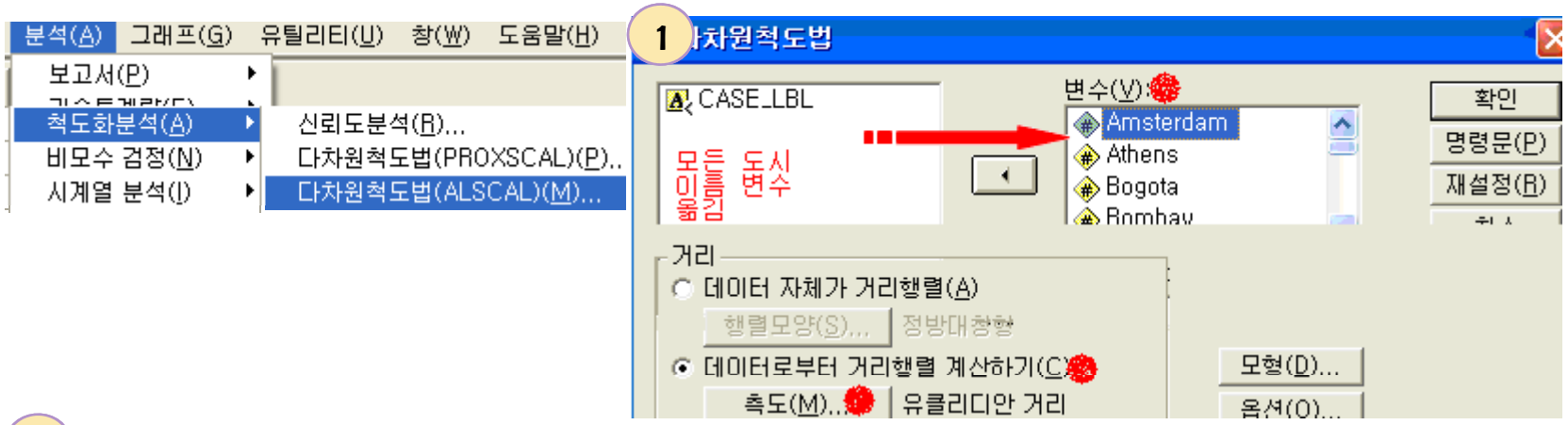
	차원	
	1	2
Amsterdam	-,561	-,036
Athens	-,263	-,106
Bogota	1,124	-,150
Bombay	,749	-,133
Brussels	-,584	-,008

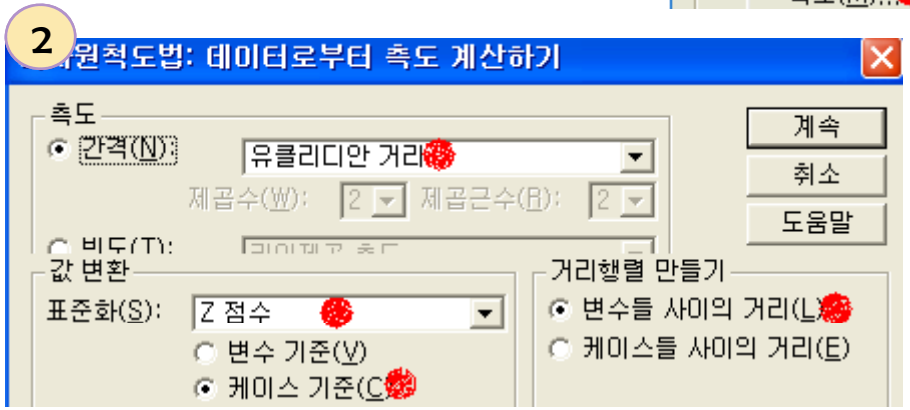
스트레스 및 적합도 측정

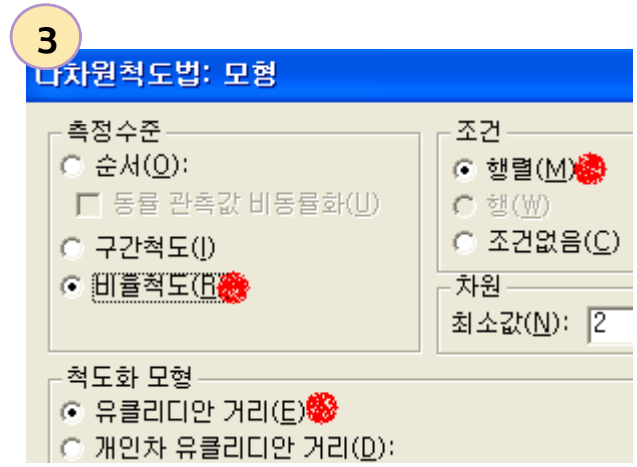
정규화된 원래 스트레스	,00083
스트레스-I	,02888[a]
스트레스-II	,04758[a]
S-스트레스	,00108[b]
설명된 산포	,99917
Tucker의 적합계수	,99958



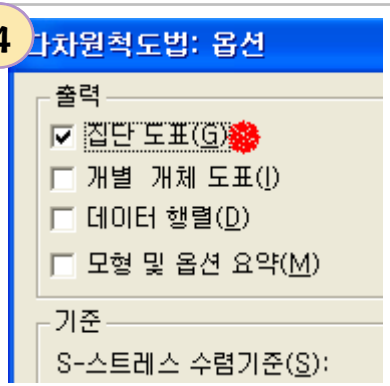
▪ 위의 예제의 경우 측정 단위가 다르므로 변수를 표준화 시켜 개체 간의 거리를 구하는 것이 바람직하다.

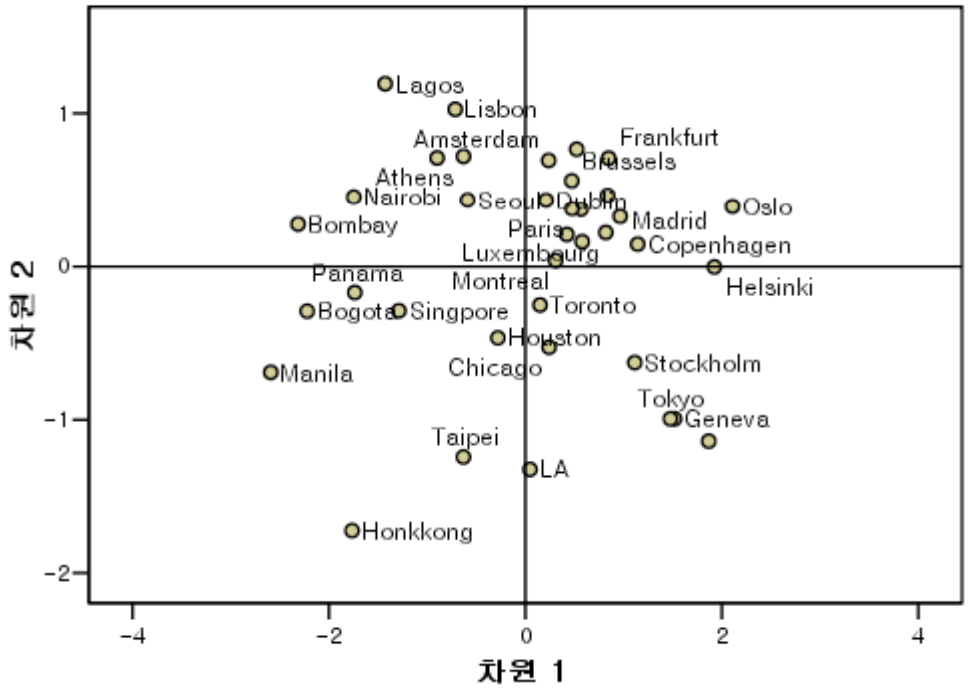






4

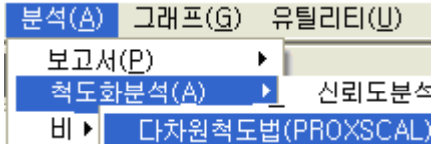

다차원척도법: 옵션
출력
집단 도표(G)
개별 개체 도표(I)
데이터 행렬(D)
모형 및 옵션 요약(M)
기준
S-스트레스 수렴기준(S):

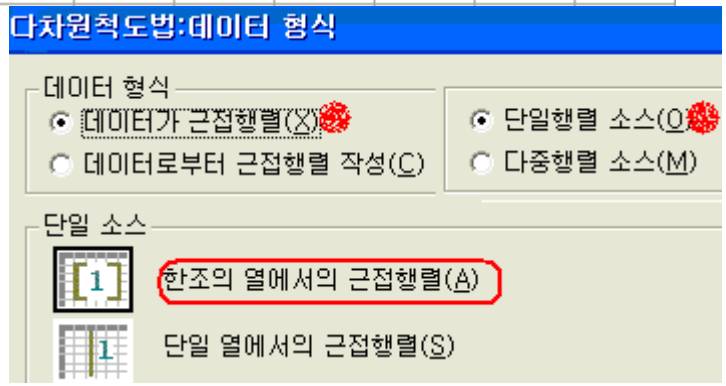

Lagos
Lisbon
Amsterdam
Frankfurt
Brussels
Athens
Nairobi
Seoul
Dublin
Oslo
Bombay
Paris
Madrid
Copenhagen
Luxembourg
Panama
Montreal
Helsinki
Bogota
Singpore
Toronto
Houston
Manila
Chicago
Stockholm
Tokyo
Geneva
Taipei
LA
Honkkong
차원 2
차원 1
1
0
-1
-2
-4
-2
0
2
4

■행렬 데이터

▪ 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

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 입력

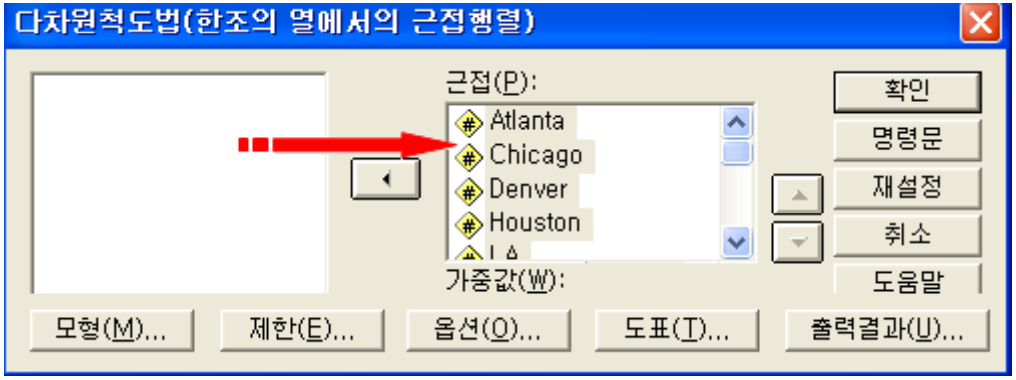
- 위와 같이 행렬의 형태로 입력한다. 첫 열의 데이터 이름과 2열부터의 변수 이름은 일치해야 한다. 이 데이터는 "거리" 변수 하나만 측정한 형태이지만 여러 항목을 측정하는 경우에도 이런 형태의 데이터 입력이 빈번히 발생한다. 데이터 입력 방법은 자료 수집을 어떻게 하였느냐에 따라 달라진다.

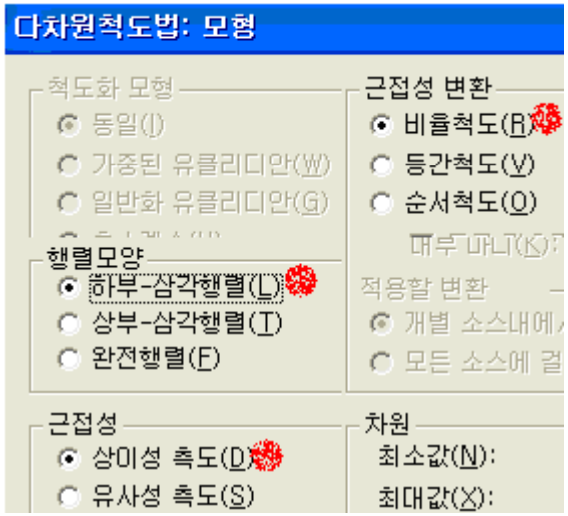
- ①아이들에게 사탕, 아이스크림, 과자, 케이크, 과일이 대한 좋아함을 10점 만점으로 측정하였다.
  - 등간 척도 다차원척도법(PROXSCAL)(P)...

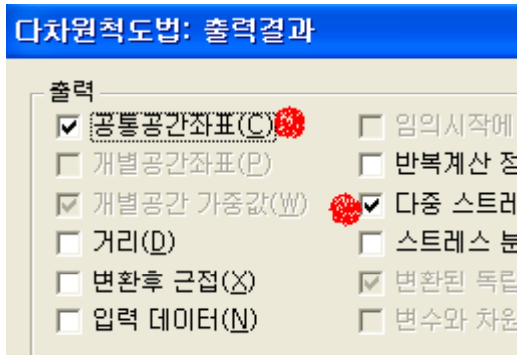
- ②아이들에게 사탕, 아이스크림, 과자, 케이크, 과일의 유사한 것부터 순위를 매겨라. 즉 사탕과 유사한 것 순위, 아이스크림과 유사한 순위,... 이렇게 하면 한 사람당 5X5 행렬이 생긴다. 그것을 사람 전체로 합하면 대칭 행렬이 된다.
  - 등간 척도 다차원척도법(ALSCAL)(M)...

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

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







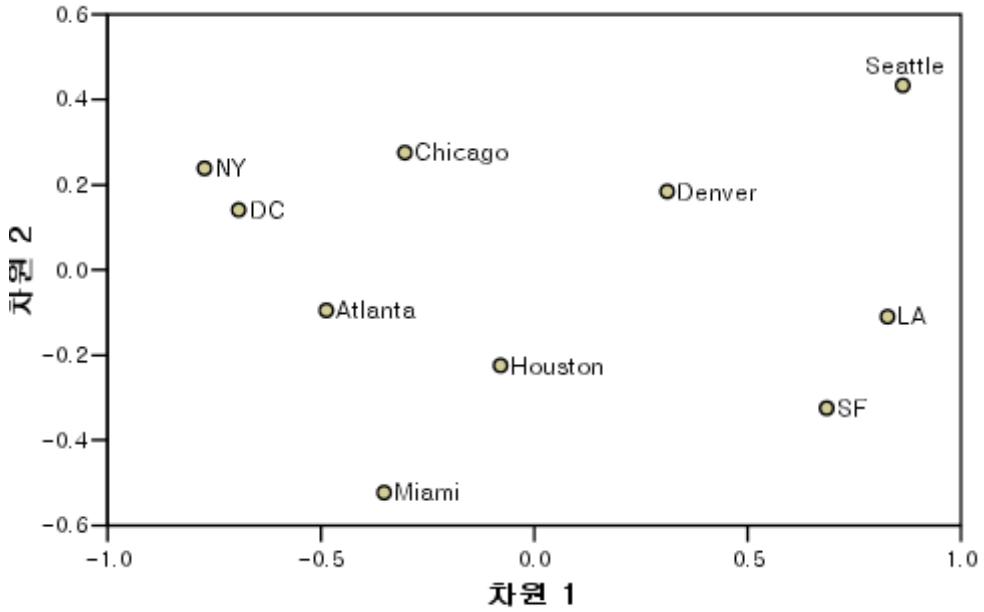
▪ 결과

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



Correspondence Analysis 개념

- 범주형 변수 범주의 유사성 표현
- 교차표(분할표)로 나타내어지는 자료의 행과 열 범주를 저차원 공간상(2차원)의 좌표로 표현하여 관계를 탐구하려는 탐색적 자료 분석 기법

기원

- 대응분석의 수리적인 기원은 1930년대 Hirshfeld의 논문『상관관계와 분할표의 연관성』
- 대응분석의 기하적인 면은 1960년대 프랑스에서 Jean-Paul Benzecri에 의해서 발전되었다.
- 일본: 1950년대 Chikio Hayashi에 의해서 수량화 제3방법으로 개발되어 발전
- 프랑스: 1960년대 Jean-Paul Benzecri가 이끄는 자료분석 모임이 다양한 분야로부터 수집된 자료를 분석하는데 대응분석 기법을 응용하고 발전

RXC 분할표

- π_{ij}: (X, Y) 결합밀도함수
- π_{i+}: (X) 주변밀도함수
- π_{+j}: (Y) 주변밀도함수

X \ Y	1	2	...	C	Total
1	π_{11}	π_{12}	...	π_{1c}	π_{1+}
2	π_{21}	π_{22}	...	π_{2c}	π_{2+}
...	...	...	...	...	...
R	π_{r1}	π_{r2}	...	π_{rc}	π_{r+}
Total	π_{+1}	π_{+2}	...	π_{+c}	π_{++}

Homogeneity (동질성)

- 각 행에 대해 열의 분포가 동일한가? $H_0 : \pi_{ij} = \pi_{kj} \ for \ j = 1,2,...,c \text{ and } k \neq i$

Independence (독립성)

- (X, Y)는 서로 독립인가? $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$

$$\text{검정통계량} = \frac{\sum_{i=1}^{k}(O_i - E_i)^2}{E_i} \sim \chi^2(k-1)$$

결과 해석

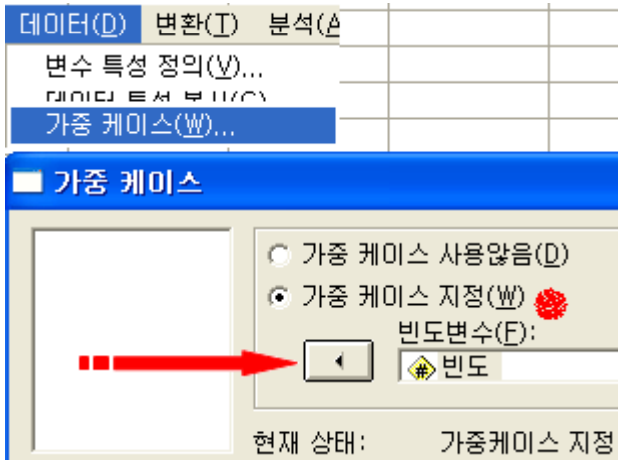
- χ^2검정 결과 p-value(유의확률) 0.05보다 적으면 두 변수 상관관계 존재
- 관계 해석: 행 페센트 혹은 열 퍼센트에 의한 차이 해석

- R×C 셀이 많아지면 퍼센트에 의한 해석이 복잡해지고 신뢰도가 떨어짐
- 행 범주, 열 범주의 유사성 정도를 표현하지 못함

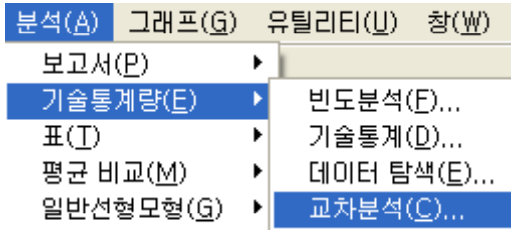
▪ 직업별 정당(a, b, c, d) 선호도의 차이가 있는지 알아보기 위하여 조사한 결과 다음 데이터를 얻었다.

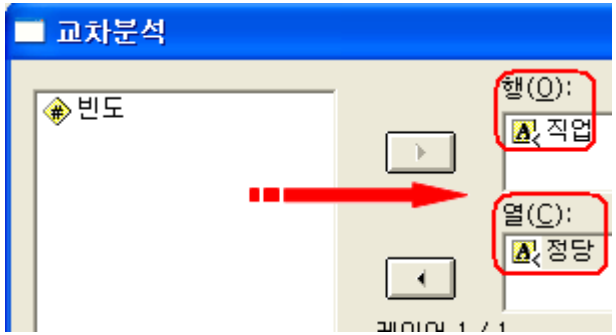
		정당			
		a	b	c	d
직업	기술직	185	240	335	130
	사무직	245	100	125	40
	전문직	105	60	70	25
	주부	40	35	75	40
	학생	40	25	30	35

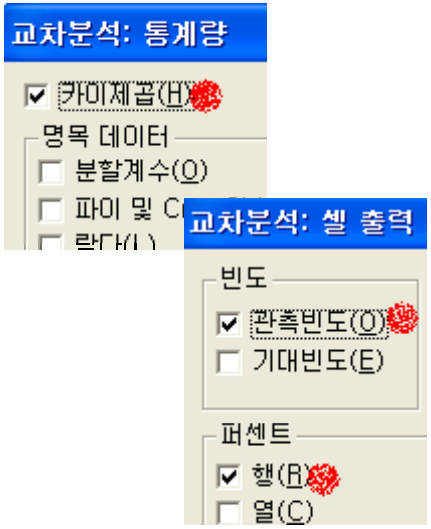
직업	정당	빈도
학생	a	40
학생	b	25
학생	c	30
학생	d	35
주부	a	40
주부	b	35
주부	c	75
주부	d	40
사무	a	245



▪ 메뉴







직업 * 정당 교차표

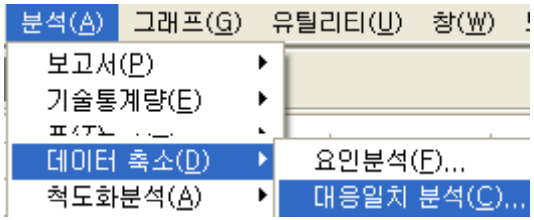
			정당				전체
			a	b	c	d	
직업	기술직	빈도	185	240	335	130	890
		직업의 %	20.8%	27.0%	37.6%	14.6%	100.0%
	사무직	빈도	245	100	125	40	510
		직업의 %	48.0%	19.6%	24.5%	7.8%	100.0%
	전문직	빈도	105	60	70	25	260
		직업의 %	40.4%	23.1%	26.9%	9.6%	100.0%
	주부	빈도	40	35	75		
		직업의 %	21.1%	18.4%	39.5%	21.1	
	학생	빈도	40	25	30		
		직업의 %	30.8%	19.2%	23.1%	26.9	
전체		빈도	615	460	635	2	
		직업의 %	31.1%	23.2%	32.1%	13.6	

카이제곱 검정

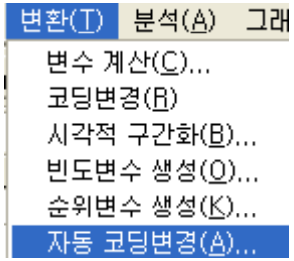
	값	자유도	점근 유의확률 (양측검정)
Pearson 카이제곱	169.120[a]	12	.000
우도비	164.952	12	.000
유효 케이스 수	1980		

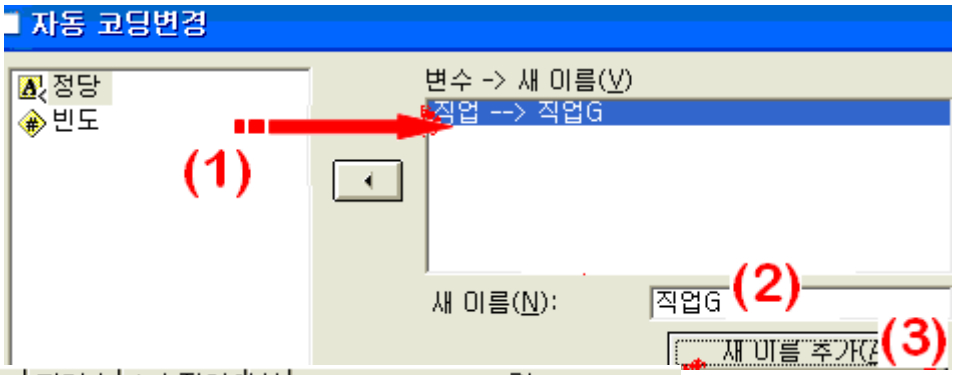
a. 0 셀 (.0%)은(는) 5보다 작은 기대 빈도를 가지는 셀입

▪ 메뉴



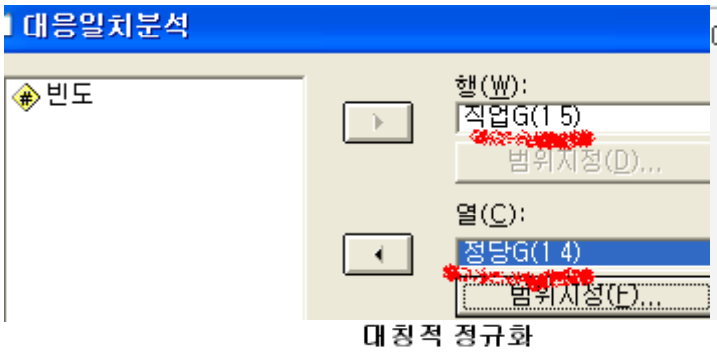
▸ 행, 열 범주 모두 숫자형이어야 한다.

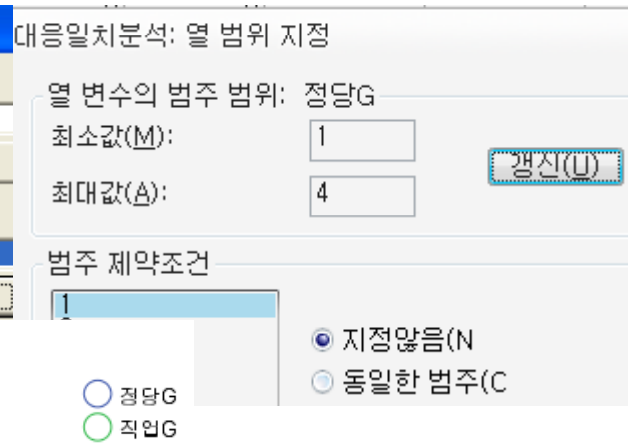


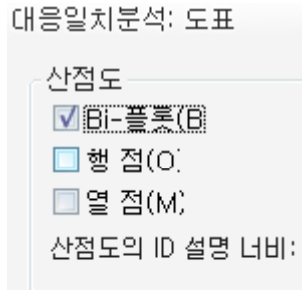


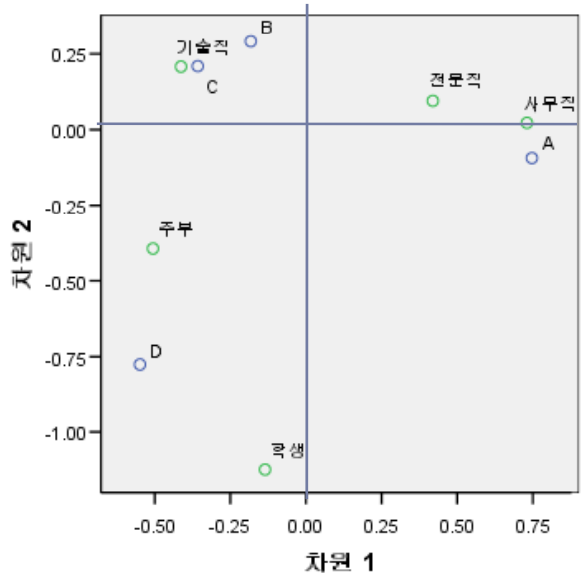
이름	유형	자리수	소수점이하설	값
직업	문자열	8	0	없음
정당	문자열	8	0	없음
빈도	숫자	4	0	없음
직업G	숫자	1	0	{1, 기술직}...
정당G	숫자	1	0	{1, a}...

▪ 메뉴 CONT.


대응일치분석
빈도
행(W):
직업G(1 5)
범위지정(D)...
열(C):
정당G(1 4)
범위지정(F)...
대칭적 정규화


대응일치분석: 열 범위 지정
열 변수의 범주 범위: 정당G
최소값(M):
1
최대값(A):
4
갱신(U)
범주 제약조건
1
지정않음(N
동일한 범주(C
정당G
직업G


대응일치분석: 도표
산점도
Bi-플롯(B
행 점(O)
열 점(M)
산점도의 ID 설명 너비:


기술직
B
C
전문직
사무직
A
주부
D
학생
0.25
0.00
-0.25
-0.50
-0.75
-1.00
차원 2
-0.50
-0.25
0.00
0.25
0.50
0.75
차원 1

- 다음 분할표는 세 부서(1, 2, 3) 직원 영어 실력 등급을 정리한 것이다. 각 셀은 빈도(도수)이다.
- 어느 부서의 영어 실력이 우수한지 알고자 한다. 적절한 분석을 하시오.

영어 \ 부서	부서1	부서2	부서3
A	78	65	68
B	22	8	30
C	20	2	7